August 2020

A Survey of Changepoint Techniques for Time Series Data

Xueheng Shi
Clemson University, shixueheng@gmail.com

A Survey of Changepoint Techniques for Time Series Data

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Mathematical Sciences

---

by
XUEHENG SHI
August 2020

Accepted by:
Dr. Colin Gallagher & Dr. Robert Lund, Committee Chair
Dr. Xin Liu
Dr. Yu-Bo Wang

Abstract

Changepoint analysis has played an important role in modern time series study. Detection of changepoints helps modelling and prediction of time series and is found in applications of many fields. This dissertation focuses on the detection of mean structure changes in correlated time series. It consists of the results of three research projects on changepoint problems: (1) the comparison of changepoint techniques; (2) autocovariance estimation of an AR(p) time series with changepoints; and (3) ℓ_1-regularization in changepoint analysis.

In chapter 2 the single changepoint techniques, or **A**t-**M**ost-**O**ne-**C**hangepoint (AMOC) are reviewed. A new AMOC test, Sum of Squared CUSUMz is developed and is shown to be the most powerful AMOC test through simulation studies on the time series with various ARMA(p, q) structures. Multiple changepoint techniques that are applicable to correlated time series are discussed in chapter 3, which includes an in-depth discussion on the wild binary segmentation. A new distance metric is also proposed in this chapter for comparing the multiple changepoint techniques. Next in the chapter 4 a Yule-Walk moment estimator based on the first order difference is proposed for autocovariance estimation of an AR(p) time series with a small number of changepoints. The last chapter simply reviews the ℓ_1 regularization and its application to changepoint analysis.

Dedication

" **To every man upon this earth, death cometh soon or late, and how can man die better, than facing fearful odds, for the ashes of his fathers, and the temples of his gods.** " – *The Lays of Ancient Rome.*

I dedicate my research and dissertation to my beloved parents. Without their selfless love and indispensable sacrifice, it would be impossible for me to complete the doctorate degree.

My parents were born several years after China's Communist Revolution. Chinese born during that period are the "ruined generation". Due to family background and political chaos, many fundamental opportunities such as the education, relocation and employment have been totally stripped from my parents. In their childhood the Three-year Great Famine (1959-1961) spread in China, about 36 million Chinese were starved. As a victim, my father lost six elder brothers, sisters and his father in the famine. My mother was lucky, my grandfather(her father) risked his life stealing food from a state-owned warehouse to support the family. Five years after the famine and when my parents were in teenage years, the Great Culture Revolution started. The massive social chaos continued for another one decade with schools closed and education cancelled. As a result, my parents became illiterates, they cannot read or write. Throughout life they can only do the dirty, unrespectable and low-paid jobs.

My parents earnestly hope my sister and me to have education so that we can live in a different way. The same thing happened to many Chinese that the parents hoped their kids to have a wealthy and stable life. Though there was a rough way, I have successfully graduated from a decent college and then went abroad to continue the graduate study. I once agonized that my parents couldn't support financially so that I could enter a dream school in the US. It's sad that I seldom carry on an in-depth conversation with them. I grow up with more experience and education, I

became aware how lonely and helpless they were. People without any education, are deaf and dumb in a modern society. They live in a complicated and fast-changing world, just like a person who cannot swim but is thrown into a whirlpool. The tragedies of my parents and their generation have dawned on me the importance of education and the meaning of life. I have been trying to stay away from the controversy and turmoil. I have been striving for a good living and learning environment. On the other hand, the poverty and misery have accompanied my parents since their birth, but they never give up the efforts and compromise on the moral standards. Their honest, humbleness and hardworking have shaped my characteristics and personalities. I left them at the age of sixteen, since then I have experienced all kinds of dilemmas, frustrations and failures. I have learned to decide independently, act calmly and response smartly.

When I was about to complete the doctorate degree and start a new chapter of my life, the COVID19 pandemic was spreading across the world in an unprecedented way. The unemployment rate soared, the political hostility towards ordinary Chinese increased and the work authorization for foreign students became precarious. I once hesitated and postponed the graduation to the fall 2020 so that I could evade the storm on the campus. But the annoying and low-paid teaching duties finally persuaded me to go ahead. I realize that I'm no longer young when I recall all the precious time that was wasted—I supposed to obtain the doctoral degree from the Georgia Institute of Technology five years ago but I failed.

I have vivid memories of my most recent trip to the United States. At that time the coronavirus started to rage my home country during the most important Lunar Spring Festival. Most Chinese suddenly entered in a panic. It became the worst Spring Festival ever in my life. Surrounding by rumors and uncertainties I immediately cancelled the vacation and booked the earliest flights back to the United States. On January 29, 2020 my parents and me left home in a hurry so that we could enter the United States before the travel ban came into effect. It was a gloomy morning with chilly winds. We sat in my brother-in-law's car silently and he played a new song to help us relax. I have been touched by the song so I translate its lyrics to English, as the

ending part of the Dedication.

Perhaps the world is always cruel and merciless.

I'm staggering forwards.

No one I can confide in, I'm lonely and silent.

My eyes have moistened.

I don't want to give up,

so I bow my head and accept all ridicules, but I look forward to the dawn.

I'm expecting to embrace the sunrise glow.

I walk forward bravely.

The dawn light finally breaks all the darkness and fears.

.....

Acknowledgments

There are many people that have earned my gratitude for their contribution to my graduate study. More specifically, I would like to thank five groups of people, without whom this dissertation would not have been possible: my doctoral advisors; my dissertation committee members; my department staffs, chair and coordinator; Clemson Palmetto Cluster Team; and my family, friends and colleagues.

First, I would like to express my sincere gratitude to my advisors **Dr. Robert Lund** and **Dr. Colin Gallagher** for their continuous support of my doctoral study and research, for their patience, motivation, and immense knowledge. Their guidance helped me in all the time of study and research. Dr. Gallagher first served as the advisor for my master thesis and he brought Dr. Lund to my thesis committee. Dr. Lund has urged me to continue the doctoral program several times, but I had been enough of the financial strains and decided to enter the job market. I became a data analyst at a research institute of Clemson University.

The job relieved my financial stress temporarily, but it was tedious. I had to analyze workers' compensation claim data every day for which I couldn't take use my statistical expertise. The employment was also unstable because the institute depended on the external contracts from AIG. I became worried about the future and felt depressed sometime. As a foreign alien I have been living under the shadow of the ruthless immigration system for near a decade. If I lost the job then I would be deprived from legal immigration status immediately. I cannot afford the cost for restarting the residency and career. After throughout considerations, I decided to resume my Ph.D via the employee benefit program. For more than one and half years I worked full-time while studied part-time. I stumbled on the qualifying exams and completed the first two subjects of statistics and stochastic process by August 2017. I will always appreciate my fellow, **Daozhou Zhu**, who is a graduate in the Department of Mathematics and comes from the same place with me in China.

Without his enormous help on the analysis exam, I would not become a doctorate candidate. My undergraduate study is engineering so I had a weak foundation on mathematical analysis. I failed twice on the analysis exam and might be dropped out from the Ph.D program again if I failed on the third attempt (I once quit Ph.D study of public policy from the Georgia Institute of Technology). During my third attempt, he sacrificed the whole winter break to help me on the analysis.

I could not have imagined having better advisors for my research. I'm indebted to my dissertation advisors, Dr. Lund and Dr. Gallagher. I started research with Dr. Gallagher in spring 2018 when he was invited to write a literature review for *Earth & Space Sciences*. He not only included me as the first author for the publication but also taught me to write an article and response reviewers' comments. Dr. Lund has believed in me like nobody else and given me endless support. He also has inspired me by his hard working and passion on research. Both have taught me fundamentals of conducting research and helped me on the technical writing. They are the models of excellent Ph.D advisors. Under their supervision, I have learned how to approach a problem, explore a solution and present the findings. Undoubtedly, I embark my academic career from their mentoring and support.

Besides my advisor, I would like to thank the rest of my dissertation committee members **Dr. Xin Liu** and **Dr. Yu-Bo Wang** for their great support and invaluable advice. During my proposal and dissertation defense, **Dr. Xin Liu** has corrected several mistakes in my work and discussed some profound questions with me. I'm also grateful to **Dr. Rebecca Killick** from Lancaster University of Great Britain, an expert of changepoint problems, for her crucial remarks and insightful comments that shaped my research. We have hundreds of email correspondences to discuss various questions on the research and collaborate on several publications.

During my study at Clemson, **Tamara Hemingway** has helped me resolve the salary dispute when I hit the rock bottom of my life in 2015. Later she adopted the aged Martin cat, who had been living around the Martin Hall for over a decade. I have taken care of Martin since 2014. However, the university was shut down during the pandemic. If Mrs. Hemingway didn't take Martin home, he would be starving. It's also worthwhile to mention that she revised the writing of the Dedication and Acknowledgment parts of my dissertation.

Finally, I would like to express my deepest gratitude to my family. This dissertation would not have been possible without their warm love, continued patience, and endless support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Changepoint Problems

Abrupt structural changes widely exist in climate, geography, economics, signal processing, bioinformatics and many other fields. These structural changes occur in the mean, variance, frequency, trend or combined.

Mean shifts play an important role in both application and theory. For example, mean shifts are a natural phenomenon in the geological exploration and can be found from the Well-logs, which are records of the physical and mineralogical characteristic of underground rocks obtained by drilling in a region of geological interest [28]: a probe is lowered into an existing well-bore by a cable and acoustical, electrical, nuclear-magnetic or thermal signals of surrounding rock types are recorded as the sonde descends. Figure 1.1 is a plot of $1,500$ time points of nuclear magnetic response. The underlying signal is piecewise constant; each segment relates to a stratum of a single rock type with constant physical properties. The discontinuities in the signal occur at times when a new stratum is first reached. Mean shift problems have been investigated by plenty of researchers. The notable methods include *binary segmentation* [38], *wild binary segmentation* [12], *MOSUM* [22], *PELT* [20], ℓ_1*-regularization* [16] and so on.

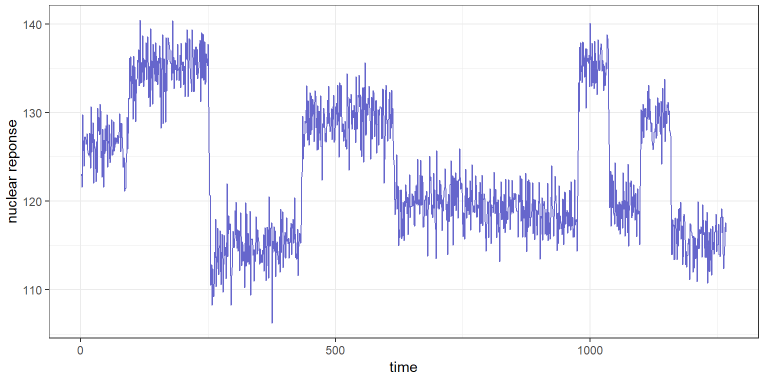


Figure 1.1: Nuclear Response Signals in Oil Drilling (well-log data)

Variance changes are less common than the mean shifts but they are still a key issue in some fields, for example, the procurement of transplant organs. Gao [14] has described an experiment on the surface temperature of a severed porcine liver. The surface temperature was constantly monitored upon the infusion of the perfusion liquid to the organ. The surface temperatures were measured every 10 minutes on a dense grid covering the whole organ for a span of 24 hours. The temperature of the organ changed in a slow fashion and maintained an overall smooth mean trend. The high oscillations in the first twelve hours reflected the resistance of the organ to the abrupt temperature change in the environment. After twelve hours, the organ started to lose the viability and the change was reflected in a sudden drop in the variance of the temperature.

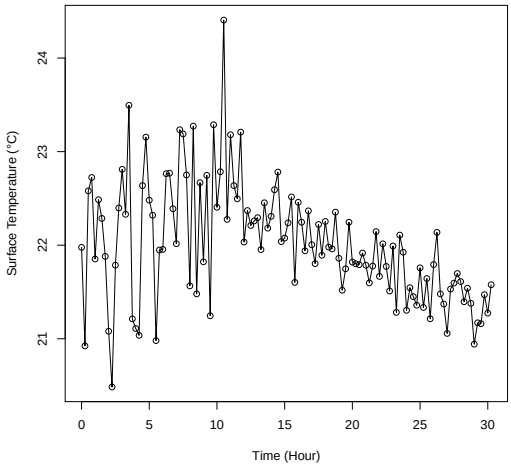
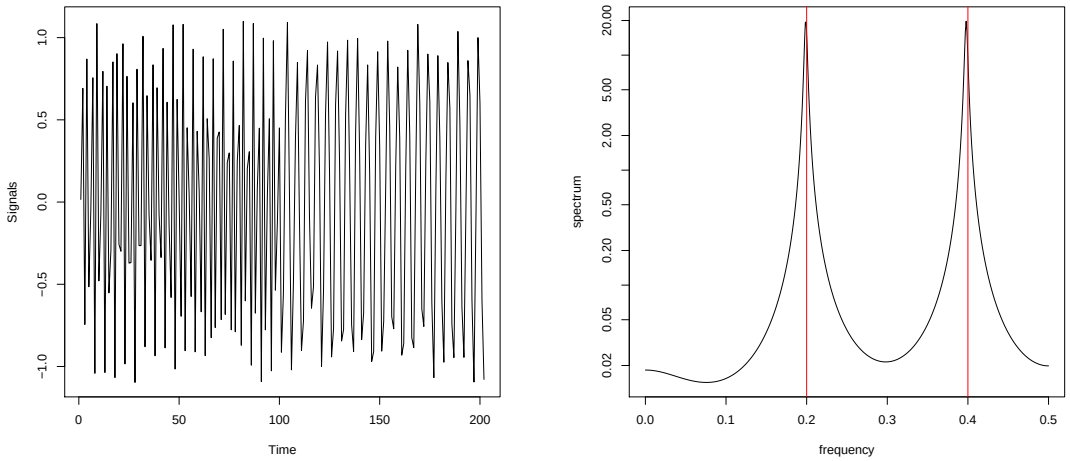


Figure 1.2: Surface Temperature of a Severed Porcine Liver

Frequency changes (note that the frequency equals to the reciprocal of the period) are popular in signal processing. An well-known example, the *Doppler effect*, is a fundamental of the modern radar and sonar systems. Doppler effect is the change in frequency of a sound or electromagnetic wave in relation to an receiver who is moving relative to the wave source. Frequency change detection often requires estimating the frequency parameters first. A review on the periodicity/frequency estimation for the unevenly spaced time series data has been written by Shi & Gallagher [43]. Frequency domain analysis has been fully developed since 1940's, due to its core application in military during World War II. The related research has established a whole new field "Spectrum Analysis".



(a) A Periodic Signal with a Frequency Change at $t = 100$

(b) Frequency Estimate via Spectrum Analysis

Figure 1.3: An Example of Frequency/Periodicity Change

In statistical analysis, changepoint detection aims to identify times when the probability distribution of a stochastic process or a time series changes. In general the changepoint analysis concerns both detecting whether or not a change has occurred, or whether several changes have occurred, and identifying the times of any such changes. Changepoint detection improves modelling and prediction of time series. A model could perform poorly or even fail when the changepoints occur in the data but was missing from the model or incorrectly specified in the model. There has been a growing demand for identifying the changepoints precisely and efficiently. Since the first published article by Page in 1954 [33], changepoint analysis becomes an increasingly important topic

of both applied and theoretical statistics. It has been actively studied with multiple changepoints, different types of data and other assumptions being considered over the past six decades.

This dissertation focuses on the detection of mean structure changes in time series data, i.e., we assume the variance and frequency structure remains unchanged. The chapters are organized as following: the introduction briefly discuss the scope of the changepoint analysis which serves as a literature review. The second chapter covers several well-known single changepoint techniques and the most powerful single changepoint technique, the SUM of Squared CUSUM on Residuals. A simulation study was conducted to disclose their strength and weakness on correlated time series. The third chapter provides an in-depth exploration on several the multiple changepoint analysis, especially the novel Wild Binary Segmentation. A simulation study was carefully designed to compare the multiple changepoint techniques. The fourth chapter focuses on the estimations for the autovariance structure of a time series in the changepoint problems. The performance of different estimates are also inspected by a simulation study. The moment estimate based on the first order difference of the time series is shown to be superior to other estimates, which is also the preliminary work for the future Gradient-Descent PELT method. The last chapter abandons the assumption of correlated structure and turns back to the sequence with i.i.d. errors so that ℓ_1 model selections are closely examined for their applications on the changepoint analysis.

Changepoints (abrupt shifts) arise in many time series due to changes in recording equipment, observers, physical moves of recording locations, or even changes in the time of day/month/year observations are recorded (this is not an exhaustive list). For example, in climatology, temperature trends computed from raw data can be misleading if homogeneity adjustments for station relocation moves and gauge changes are not *a priori* made to the record. Lu & Lund [25] gave an example where trend conclusions are spectacularly spurious when changepoint information is neglected. Multiple changepoints are also frequently encountered; for example, in climatology, United States climate stations average about six station moves and/or gauge changes per century of operation [31].

This dissertation is intended to guide the researcher on the best changepoint techniques to use in common scenarios. Assumptions are crucial in changepoint analyses and can significantly alter conclusions; here, issues of correlation will take center stage. It is known that changepoint inferences made from positively correlated series can be spurious if correlation is not taken into account. Even lag one correlations as small as 0.25 can have deleterious consequences on changepoint conclusions

[27].

This dissertation's primary contribution is twofold: 1) extend/modify many of the popular changepoint methods for i.i.d. data to correlated settings, 2) compare these methods against each other, 3) autocovariance estimation based on the first order difference of the series in the changepoint problems. Much of my work lies with developing methods that puts all techniques, to the best extent possible, on the same footing in time series settings. For example, we will see that single changpoint tests generally work best when applied to estimated versions of the time series one-step-ahead prediction residuals, computed under a null hypothesis of no changepoints. Because of this, tests that handle one-step-ahead prediction residuals need to be developed. The comparative aspect of the paper is a second central contribution — and there is much to compare. For example, in detecting a single changepoint, one could use a statistic that asymptotically converges to an extreme value distribution when maxed over all admissible changepoint locations, or "crop some admissible changepoint locations" near the data boundaries and scale to a supremum of a limiting Gaussian process (which are typically related to Brownian Bridges). In addition to comparing different statistics via Type I errors and powers, the dissertation also compares different asymptotic scaling methods.

With these lofty objectives, some concessions are necessary. Foremost, this paper examines mean shift changepoints only; that is, while series mean levels are allowed to abruptly shift, the variances and correlations of the series are held stationary in time. Changepoints can also occur in variances (volatilities) (See [18]), in the series' correlation structures (see [9]), or even in the marginal distribution of the series. For example, one changepoint test for marginal distributions is [13], where changepoint methods for shifts in daily precipitation series are developed. Here, precipitation has a marginal distribution with a point mass at zero for dry days and a conditional density over $(0, \infty)$ for rainfall amounts on wet days. Secondarily, the simulation results reported here are primarily for Gaussian series. For strictly stationary series with a unimodal marginal distribution, results will be similar in spirit to those obtained here. The same claim cannot be made for series whose marginal distribution is exotic — say a count distribution that is supported on the skewed support set $\{1, 2, 3, 25, 500\}$. Here, means and variances would insufficiently describe the problem's statistical structure. Finally, although ℓ_1-regularizations are worthy to consider, preliminary work shows that ℓ_1 methods perform poorly without significant alterations [40]. The alterations needed to make ℓ_1 methods competitive in changepoint detection are extensive and will be dealt with in a distinct

manuscript.

Academic changepoint research commenced with the single changepoint case for independent and identically distributed (i.i.d.) data in [33]. The changepoint subject is now vast, with thousands of papers devoted to the topic. Worth citing is [29], where asymptotic theory for a single changepoint is developed. Single time series changepoint contributions include [8, 1, 36, 19, 37] and the references within — this is by no means a complete list. Multiple changepoint techniques for correlated series are reviewed in [32], although the spirit of their review is different than ours here. Binary and wild binary segmentation techniques — methods that recursively apply a single changepoint test to sub-segments of the series to identify all changepoints — are explained and developed in [12]. General time series changepoint techniques for a regression response more complicated than simple mean shifts are considered in [30, 37].

The rest of this dissertation proceeds as follows. The next chapter overviews single changepoint detection methods, typically termed at most one changepoint (AMOC) tests. Here, a variety of test statistics and their scalings are reviewed and tuned to the time series setting. We then compare AMOC detectors in a simulation study. Here, a technique that uses the argument of a CUSUM-type statistic to identify where the changepoint occurs, but uses the sum of the squared CUSUM statistics over all locations to assess whether a changepoint exists, is shown to have a good Type I error and superior detection power. We also show that AMOC tests typically work better when applied to the time series one-step-ahead residuals, which are always uncorrelated (independent for Gaussian series). The alternative modifies statistics to account for the correlation — see Theorem 1 below. The multiple changepoint case is more nebulous; here, some techniques work well for some multiple mean shift configurations and poorly on others. Of note here is the development of a new distance between two changepoint configurations that allows us to compare methods. Lastly, the autocovariance estimation in changepoint problems and ℓ_1 regularization approaches are briefly discussed.

Readers, please be aware that this dissertation was written during the COVID19 pandemic. The author could not access the Writing Center at Clemson University to seek help on English writing which is crucial for an non-native speaker. Though the author has done his best, due to the lack of time and heavy research work, the dissertation may contain grammar and typo errors, some descriptions may be inaccurate and inappropriate. Please refer to the author's publications.

Chapter 2

Single Changepoint Techniques

2.1 Single Changepoint Problem

Changepoint techniques were first developed for the simpler At-Most-One-Changepoint(AMOC) scenario. AMOC approaches will be first considered before multiple changepoints are studied in chapter 3.

Let $\{X_t\}_{t=1}^N$ be the observed time series and $\gamma(h) = \text{Cov}(X_{t+h}, X_t)$ be the lag h autocovariance of the series. While the first moment of $\{X_t\}$ may shift, the second moment is assumed stationary in time. An AMOC model, with the changepoint occurring at the unknown time $k+1$, is

$$X_t = \begin{cases} \mu + \epsilon_t, & \text{for } 1 \leq t \leq k, \\ \mu + \Delta + \epsilon_t, & \text{for } k+1 \leq t \leq N, \end{cases} \tag{2.1}$$

where μ is unknown, Δ is the magnitude of mean shift at time $k+1$, and $\{\epsilon_t\}$ is a stationary time series with zero mean and lag h covariance $\gamma(h)$. A hypothesis test for this scenario is:

$$H_0 : \Delta = 0 \quad \text{versus} \quad H_1 : \Delta \neq 0 \quad \text{for some } k \in \{1, \ldots, N-1\}. \tag{2.2}$$

Two classic AMOC techniques, cumulative sum (CUSUM) statistics and likelihood ratio (LR) tests, are now considered.

2.2 CUSUM Tests

The CUSUM method was first introduced by [33] and compares sample means before and after each admissible changepoint time k. Scaling differences between sample means before and after time k, $\bar{X}_k = k^{-1}\sum_{t=1}^{k} X_t$ and $\bar{X}_k^* = (N-k)^{-1}\sum_{t=k+1}^{N} X_t$, to a nondegenerate probability distribution led to the statistic

$$\text{CUSUM}_X(k) := \frac{1}{\sqrt{N}}\left[\sum_{t=1}^{k} X_t - \frac{k}{N}\sum_{t=k+1}^{N} X_t\right]. \tag{2.3}$$

The location of the largest absolute CUSUM index is estimated as the location of the changepoint time and

$$\max_{1<k\leq N} |\text{CUSUM}_X(k)|. \tag{2.4}$$

is taken as the test statistic. To quantify the asymptotic distribution of the statistic in 2.3, we assume that $\{\epsilon_t\}$ has the usual causal linear representation

$$\epsilon_t = \sum_{i=0}^{\infty} \psi_i Z_{t-i}, \quad t \in \mathbb{Z}. \tag{2.5}$$

Here, $\{Z_t\}_{t\in\mathbb{Z}}$ is IID with zero mean, variance σ^2, finite fourth moment $E[Z_t^4]$, and $\sum_{j=0}^{\infty}|\psi_j| < \infty$. All causal autoregressive moving-average (ARMA) models admit such a representation.

The asymptotic distribution of the statistic in (2.4), under the null hypothesis of no changepoints, is well known [29, 8]. Let $\{B(t), t \in [0,1]\}$ be a standard Brownian Bridge process obeying $B(t) = W(t) - tW(1)$, where $\{W(t), t \geq 0\}$ is a standard Wiener process. Define the long-run process variance as

$$\eta^2 = \lim_{n\to\infty} \frac{1}{n}\text{Var}\left(\sum_{t=1}^{n} \epsilon_t\right). \tag{2.6}$$

The following result is known from [8].

Theorem 1. If $\{X_t\}$ is assumed to follow (2.1) and $\hat{\eta}^2$ is a null hypothesis based consistent estimator of η^2, then under H_0,

$$\frac{1}{\hat{\eta}} \max_{1<k\leq N} |\text{CUSUM}_X(k)| \xrightarrow{\mathcal{D}} \sup_{t\in[0,1]} |B(t)|. \tag{2.7}$$

Under H_0, the asymptotic distribution of the CUSUM statistic follows the probability law

(see section 6.10 in [34])

$$\mathbb{P}\left[\sup_{t\in[0,1]}|B(t)|>x\right]=2\sum_{n=1}^{\infty}(-1)^{n+1}e^{-2n^2x^2},\quad x>0. \tag{2.8}$$

Estimation of η^2, which is related to the spectral density of $\{X_t\}$, is often difficult; see Chapter 2 of [44].

2.3 CUSUM Tests for ARMA Residuals

We now consider the case where $\{\epsilon_t\}$ is correlated, perhaps a causal and invertible ARMA(p,q) series obeying

$$\epsilon_t-\phi_1\epsilon_{t-1}-\cdots-\phi_p\epsilon_{t-p}=Z_t+\theta_1Z_{t-1}+\cdots+\theta_qZ_{t-q},\quad t\in\mathbb{Z}, \tag{2.9}$$

where $\{Z_t\}$ is IID with zero mean, variance σ^2, and a finite fourth moment.

Strong correlation often degrades CUSUM performance [36]; that is, convergence to the limit law happens faster for independent data than for positively correlated data. As such, it is often beneficial to decorrelate heavily dependent data before applying CUSUM methods.

The ARMA one-step-ahead prediction residuals are computed by mimicking (2.9):

$$\hat{Z}_t=X_t-\hat{\phi}_1X_{t-1}-\cdots-\hat{\phi}_pX_{t-p}-\hat{\theta}_1\hat{Z}_{t-1}-\cdots-\hat{\theta}_q\hat{Z}_{t-q}, \tag{2.10}$$

where $X_t=\hat{Z}_t=0$ for any $t<0$. The estimator $\hat{\sigma}^2=N^{-1}\sum_{t=1}^{N}\hat{Z}_t^2$ is used. The residual CUSUM statistic at time k is

$$\text{CUSUM}_Z(k)=\frac{1}{\sqrt{N}}\left(\sum_{t=1}^{k}\hat{Z}_t-\frac{k}{N}\sum_{t=k+1}^{N}\hat{Z}_t\right). \tag{2.11}$$

The following result is taken from [36].

Theorem 2. *Suppose that $\{\epsilon_t\}$ is a causal and invertible ARMA series with IID $\{Z_t\}$ with $E[Z_t^4]<\infty$, and let $\{\hat{Z}_t\}$ be the estimated one-step-ahead prediction residuals in (2.10). Then under the null hypothesis of no changepoints,*

$$\frac{1}{\hat{\sigma}}\max_{1<k\le N}|CUSUM_Z(k)|-\frac{1}{\hat{\eta}}\max_{1<k\le N}|CUSUM_X(k)|=o_p(1), \tag{2.12}$$

when all ARMA parameters and τ^2 are estimated in a $\sqrt{N}$-consistent manner. It hence follows that

$$\frac{1}{\hat{\sigma}} \max_{1<k\leq N} |CUSUM_Z(k)| \xrightarrow{\mathcal{D}} \sup_{0\leq t\leq 1} |B(t)|. \tag{2.13}$$

Critical values for CUSUM statistics can be obtained via simulation; listed here are some common percentiles.

Table 2.1: Critical Values for CUSUM Statistics

Percentile	Critical Value
90.0%	1.224
95.0%	1.358
97.5%	1.480
99.0%	1.628

2.4 Cropped CUSUM Tests

CUSUM tests have relatively poor detection power when the changepoint occurs near the boundaries (times 1 or N). Conversely, false detection is more likely to be signaled at boundary locations. This is expected since few observations lie between the changepoint and the boundary and estimation of a segment mean may be less precise. Mathematically, [8] address this problem by applying a weight function $w(\cdot)$, denoted by $w(t)$ at time $t = k/N$.

For example, with $w(t) = \sqrt{t(1-t)}$ and

$$\lambda_X(k) = \frac{\text{CUSUM}_X^2(k)}{\frac{k}{N}(1-\frac{k}{N})} \text{ and } \lambda_Z(k) = \frac{\text{CUSUM}_Z^2(k)}{\frac{k}{N}(1-\frac{k}{N})}. \tag{2.14}$$

Theorem ??? from [36], states the following.

Theorem 3. *Given $0 < \ell < h < 1$ and suppose that $\hat{\sigma}^2$ and $\hat{\tau}^2$ are $\sqrt{N}$-consistent estimates of σ^2 and τ^2 respectively. Under H_0,*

$$\frac{1}{\hat{\eta}^2} \max_{\ell\leq k/N\leq h} \lambda_X(k) \xrightarrow{\mathcal{D}} \sup_{\ell<t<h} \frac{B^2(t)}{t(1-t)}, \tag{2.15}$$

and

$$\frac{1}{\hat{\sigma}^2} \max_{\ell \le k/N \le h} \lambda_z(k) \xrightarrow{\mathcal{D}} \sup_{\ell < t < h} \frac{B^2(t)}{t(1-t)}. \tag{2.16}$$

One can approximate p-values for cropped CUSUM tests via

$$\mathbb{P}\left[\sup_{\ell \le t \le h} \frac{B^2(t)}{t(1-t)} > x\right] \approx \sqrt{\frac{xe^{-x}}{2\pi}} \left[\left(1 - \frac{1}{x}\right) \log\left(\frac{(1-\ell)h}{\ell(1-h)}\right) + \frac{4}{x}\right]. \tag{2.17}$$

2.5 Sum of Squared CUSUM Statistics

Relatively recently, [21] proposed summing the squared CUSUM statistics over all time indices. As we will soon see, this test has larger detection power than other AMOC approaches. As before, the time with the largest absolute CUSUM statistic is estimated as the changepoint time.

Note that (2.11) can be written as

$$\text{CUSUM}_Z(k) = \frac{k}{N}\left(1 - \frac{k}{N}\right)\sqrt{N} \cdot \bar{Z}_k - \frac{k}{N}\left(1 - \frac{k}{N}\right)\sqrt{N} \cdot \bar{Z}^*_k, \tag{2.18}$$

where $\bar{Z}_k = \frac{1}{k}\sum_{t=1}^{k} \hat{Z}_t$ and $\bar{Z}^*_k = \frac{1}{N-k}\sum_{t=k+1}^{N} \hat{Z}_t$. Under the null hypothesis of no changepoints, the central limit theorem provides

$$\frac{\text{CUSUM}_Z(k)}{\hat{\sigma}} \xrightarrow{\mathcal{D}} \mathbf{N}\left(0, \frac{k}{N}\left(1 - \frac{k}{N}\right)\right). \tag{2.19}$$

This holds for i.i.d. data and asymptotically for one-step-ahead prediction residuals when σ^2 and all ARMA parameters are estimated $\sqrt{N}$-consistently.

Let $t = k/N$. By the functional central limit theorem (See Section 8 of [3]), the process-based convergence

$$\left\{\frac{\text{CUSUM}_Z(k/N)}{\hat{\sigma}}\right\} \xrightarrow{\mathcal{D}} \{B(t)\}_{t=0}^{t=1} \tag{2.20}$$

can be shown to hold weakly, where $\{B(t)\}_{t=0}^{t=1}$ is a Brownian bridge. On the sample paths of Brownian bridges, sum of squared paths converge to integrals of squared paths; hence, application

of the Continuous Mapping Theorem provides

$$\text{SCUSUM}_Z(k) = \frac{1}{N}\sum_{k=1}^{N}\left[\frac{\text{CUSUM}_Z(k/N)}{\hat{\sigma}}\right]^2 \xrightarrow{\mathcal{D}} \int_0^1 B^2(t)\,dt. \tag{2.21}$$

The distribution of $\int_0^1 B(t)^2 dt$ was investigated in [45] which obeys the following distribution function:

$$F(t \leq \lambda) = 1 - \frac{2}{\pi}\sum_{k=1}^{\infty}\int_{(2k-1)\pi}^{2k\pi} \frac{\exp(-\frac{x^2\lambda}{2})}{\sqrt{-x\sin x}}\,dx. \tag{2.22}$$

A simulation was conducted to obtain critical values for the Sum of Squared CUSUM test; these are reported below for convenience. A plot of the distribution of the integral of the squared Brownian Bridge is also attached.

Table 2.2: Critical Values for Sum of Squared CUSUM Statistics

Percentile	Critical Value
90.0%	0.3473046
95.0%	0.4613744
97.5%	0.5806168
99.0%	0.7434348

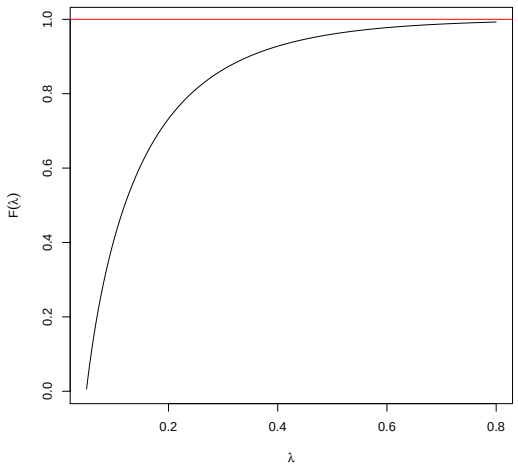


Figure 2.1: The Cumulative Distribution Function of the Integral of the Squared Brownian Bridge

2.6 Likelihood Ratio Tests

Likelihood ratio tests (LRT) split the series into the segments $\{X_1, \cdots, X_k\}$ and $\{X_{k+1}, \cdots, X_N\}$ and compute alternative hypothesis likelihoods by allowing the two segments to have distinct means, say μ_1 and $\mu_2 = \mu_1 + \Delta$, respectively. The likelihood $L_k(\mu_1, \mu_2)$ under H_1 is then compared to the likelihood $L_0(\mu_0)$ under H_0 (μ_0 denotes the common mean of the series). Computation of the likelihoods, in general, allows for correlation in the series. This is realitively easy when model errors are from a Gaussian ARMA process [5], but may be considerably harder under other distributional assumptions. Any correlation parameters — for example, those arising in the ARMA model — are nuisance parameters in so far as the changepoint is concerned.

The LRT statistic for a changepoint at time k has the general form

$$\Lambda = \max_{1<k\leq N} \left[\frac{L_0(\hat{\mu}_0)}{L_k(\hat{\mu}_1, \widehat{\mu_2})} \right], \tag{2.23}$$

where $\hat{\mu}_0$ is the maximum likelihood estimator (MLE) for the mean of $\{X_t\}$ under H_0, and $\hat{\mu}_1$ and $\hat{\mu}_2$ are the MLEs for the means of the two segments under H_1.

Two different scalings of LRTs have been developed that converge to extreme value distributions and supremums of Gaussian processes, respectively. Uncropped versions of these statistics are [19]

$$U = \max_{1<k\leq N} \left(-2\log(\Lambda_k)\right), \quad \text{where } \Lambda_k = \left(\frac{\hat{\sigma}_k^2}{\hat{\sigma}_{H_0}^2} \right)^{\frac{N}{2}} \tag{2.24}$$

$$W_U = \sqrt{2U\log\log(N)} - \left[2\log\log(N) + \frac{1}{2}\log\log\log(N) - \frac{1}{2}\log\pi \right]. \tag{2.25}$$

Here, the following clarifications are made. Under H_0, the one-step-ahead ARMA prediction residuals $\{\hat{Z}_t\}$ in (2.10) are computed and the white noise variance of $\{Z_t\}$ is estimated under H_0 as $\hat{\sigma}_{H_0}^2 = N^{-1}\sum_{t=1}^{N} \hat{Z}_t^{\,2}$. Under H_1, a different estimated piecewise mean for $\{X_t\}$ exists for each admissible changepoint time k. Calling time k estimates $\mu_{1,k}, \mu_{2,k}, \phi_k, \Delta_k, \sigma_k^2$, etc., one first computes one-step-ahead prediction residuals for $\{X_t\}$. For concreteness, the sample mean $\hat{\mu}_{t,k} = \widehat{E[X_t]}$ is

$$\hat{\mu}_{t,k} = \begin{cases} \hat{\mu}_{1,k}, & 1 \leq t \leq k \\ \hat{\mu}_{1,k} + \hat{\Delta}_k, & k+1 \leq t \leq N, \end{cases} . \tag{2.26}$$

For example, in the case of a Gaussian AR(1) series with a changepoint under H_1 at time k, the estimated one-step-ahead prediction residual at time t is $[X_t - \hat{\mu}_{t,k}] - \hat{\phi}_k[X_{t-1} - \hat{\mu}_{t-1,k}]$ and

$$\hat{\sigma}_k^2 = \frac{1}{N}\sum_{t=2}^{N}[(X_t - \hat{\mu}_{t,k}) - \hat{\phi}_k(X_{t-1} - \hat{\mu}_{t-1,k})]^2. \tag{2.27}$$

For convenience, sample means are used in place of the true mean parameter MLEs as there is usually no practical difference in them (for Gaussian cases); however, MLEs are much harder to compute under correlation. This implies that

$$\hat{\mu}_{1,k} = \frac{1}{k}\sum_{t=1}^{k} X_t, \qquad \hat{\mu}_{2,k} = \hat{\mu}_{1,k} + \hat{\Delta} = \frac{1}{N-k}\sum_{t=k+1}^{N} X_t. \tag{2.28}$$

The statistic W_U can be asymptotically scaled to a Gumbel type distribution under H_0:

$$\lim_{N\to\infty} \mathbb{P}(W_U \leq x) = \exp(-2\exp(-x)), \qquad -\infty < x < \infty. \tag{2.29}$$

Here, H_0 is rejected when W_U is too large to be explained by the distribution in (2.29).

[36] shows that $-2\log(\Lambda_k)$ is related to the CUSUM $\lambda_Z(k)$ statistic through

$$\max_{\ell \leq k/N \leq h}\{-2\log(\Lambda_k)\} - \frac{1}{\hat{\sigma}^2}\max_{\ell \leq k/N \leq h}\lambda_Z(k) = o_p(1). \tag{2.30}$$

Thus, if $\ell \downarrow 0$ and $h \uparrow 1$, $\text{CUSUM}_Z(k)$ and LRTs are linked by

$$T = \frac{1}{\hat{\sigma}^2}\max_{\ell \leq k/N \leq h}\lambda_Z(k), \tag{2.31}$$

$$W_T = \sqrt{2T\log\log(N)} - \left[2\log\log(N) + \frac{1}{2}\log\log\log(N) - \frac{1}{2}\log\pi\right]. \tag{2.32}$$

As $N \to \infty$, W_U converges to the Gumbel distribution in (2.29). There is no need to crop boundaries here as extreme value scalings allow all admissible changepoint times to be considered.

Cropped LRTs simply truncate admissible times at the boundaries; for example,

$$U_{\text{crop}} = \max_{\ell \leq k/N \leq h}(-2\log(\Lambda_k)). \tag{2.33}$$

Connections exist between U_{crop} and $\lambda_X(k)$:

$$\max_{\ell \le k/N \le h} (-2 \log \Lambda_k) - \frac{1}{\hat{\eta}_2^2} \max_{\ell \le k/N \le h} (\lambda_X(k)) = o_p(1). \tag{2.34}$$

This identifies U_{crop}'s asymptotic null hypothesis distribution.

Theorem 4. *Under H_0, the cropped LRT statistic obeys*

$$U_{\text{crop}} = \max_{\ell \le k/N \le h} (-2 \log(\Lambda_k)) \xrightarrow{\mathcal{D}} \sup_{\ell \le t \le h} \frac{B^2(t)}{t(1-t)}. \tag{2.35}$$

2.7 AMOC Simulations

This section investigates the finite sample performance of the Section 2.1 tests through simulation. We are interested in the impact of autocorrelation on the tests. To start, first order Gaussian autoregressions (AR(1)) are examined; here, the lag-one correlation equals ϕ and the entire correlation structure is quantified by a single parameter. More complex correlation structures are considered thereafter. Methods with good false detection rates when no changepoints are present are sought, regardless of the degree of correlation. Desirable tests have reasonable (non-inflated) false detection rates when no changepoints exist, and large detection powers when a changepoint is present. We first explore the impact of autocorrelation on the Type I error, finding that some tests break down badly under certain scenarios. We then move to comparing detection powers of the tests with good Type I errors when a changepoint exists.

Table 4 lists empirical false detection rates (Type I Errors) based on 10,000 independent simulated series in a 95% AMOC test for several values of ϕ and N. Table 4 bolds false detection rates that exceed the nominal 0.05 level by two standard errors, which is binomally based — $\sqrt{\hat{p}(1-\hat{p})/N}$, where $\hat{p}$ is the proportion of runs that reject H_0. The first two columns of Table 4 show that tests based on the raw data (as opposed to one-step-ahead prediction residuals) have a large Type I error in the presence of strong negative autocorrelation (one might argue that this case is seldom encountered in practice). This was also noted in [36], where more generally, tests based on ARMA one-step-ahead prediction residuals were shown to be preferable to analogous tests for the raw data, but adjusted for correlation (for example, the η in Theorem 1 is a factor that adjusts for correlation). The next three columns show tests, all based on ARMA residuals, where the Type I error is non-

inflated across all correlations: the maximum of CUSUM_Z in (2.13), the maximum of λ_Z in (2.16), and the sum of squared SCUSUM_Z in (2.21), respectively. The last three columns of Table 4 display results for three LRT tests. While the test in (2.24) outperforms the other LRTs, Figure 2.2 plots Type I error as a function of ϕ for the three ARMA residuals and the test in (2.24). Three of these tests control false positive rates well, regardless of the degree of correlation, while the LRT test has very conservative Type I errors for most ϕ, but then becomes very inflated as the correlation approaches unity (a case that does arise in practice). It does not appear that LRT tests work well.

We proceed by considering the power of the four tests shown in Figure 2.3. In general, the detection power of AMOC tests depend on the degree of correlation, the size of the mean shift, and the location of the changepoint time [37]. In Figure 2.3, empirical powers based on 10,000 independent Gaussian simulated series with sample size $N = 1,000$ are plotted as a function of ϕ when the mean shift lies in the centre of the series (time 500). We first simulate a small mean shift, $\Delta = 0.15$, to demonstrate the drastic effects of autocorrelation. The CUSUM_Z and SCCUSUM tests are more powerful than the others. Note also that SCUSUM_Z has higher power than CUSUM_Z for each ϕ considered. Additional simulations (not shown) duplicate this conclusion for other sample sizes. While the LRT had the highest empirical power when $\phi = .95$, this test also has a Type I error far exceeding 0.05; hence, such power is not indicative of better overall performance. The sum of squared CUSUM statistic is the overall winner so far.

We now study the effect of the changepoint time location on results. Simulation specifications are as in the above paragraph. In general, detection powers are largest when the changepoint occurs near the center of the record; power decreases as the changepoint time moves towards a boundary. This is seen in Figure 2.5, which plots empirical powers as a function of the changepoint location. The test based on SCUSUM_Z appears to be best overall. However, the LRT test is preferable when the changepoint occurs near the beginning of the record and the weighted/cropped CUSUM of the residuals is more powerful than the sum of squared CUSUM test when the changepoint is near the end of the data sequence. Again, the sum of squared CUSUM test is the overall winner.

Some slight asymmetry in the empirical powers may be noted. This was investigated and traced to the fact that the distribution of $\hat{\tau}$ is asymmetric in time. This, in turn, traces to the fact that time 1 cannot be a changepoint time while time N can be.

To study different correlation structures, other ARMA models are considered. Table 3 shows

analogous type I errors and detection powers (when the mean shift is placed in the centre of the record) for several other low-order ARMA models. The results here are much in the spirit of the above, with the sum of squared CUSUM statistic being the overall winner. The reader is encouraged to examine the supplementary material for additional simulations. Our overall recommendations are now summarized:

- Extreme value and LRTs can have poor Type I error and/or detection power and should not be trusted without extensive scrutiny.
- CUSUM-based tests applied to estimated versions of the one-step-ahead prediction residuals work better than tests that try to adjust classical statistics for independent series via η.
- The SCUSUM$_Z$ statistic in (2.21) generally outperforms the other tests, especially when applied to one-step-ahead prediction residuals. It has a stable false detection rate near the nominal level, regardless of the degree of correlation, and has higher empirical power than the other tests in nearly all simulated scenarios. As such, we recommend using the SCUSUM$_Z$ statistic (applied to one-step-ahead prediction residuals) in the AMOC mean shift setting for correlated data.

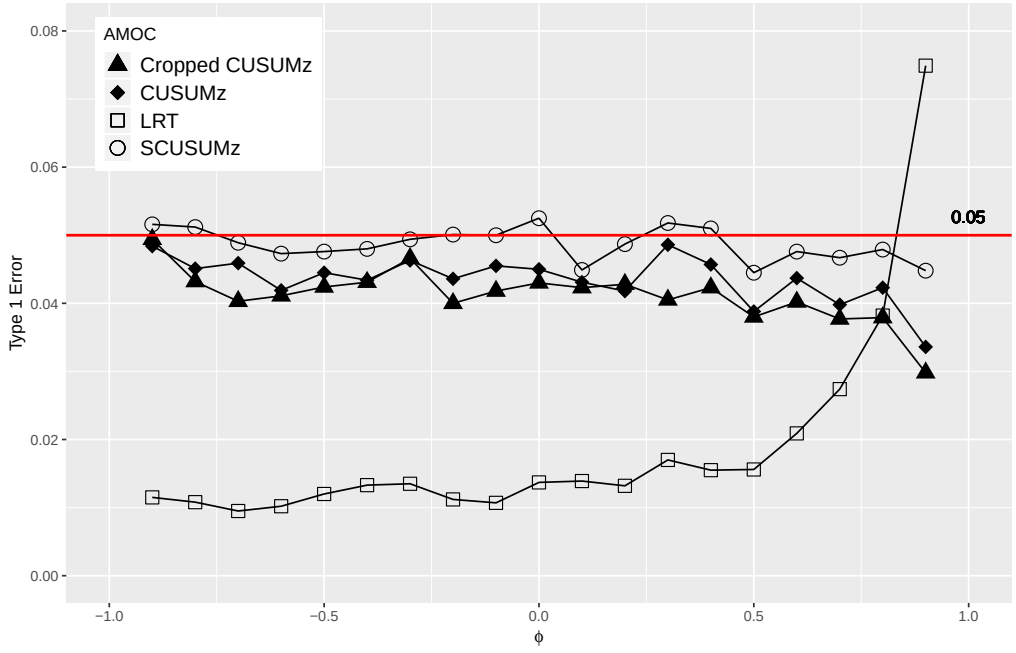


Figure 2.2: Graph of Type I Errors for an AR(1) series with Different ϕ when $N = 1000$.

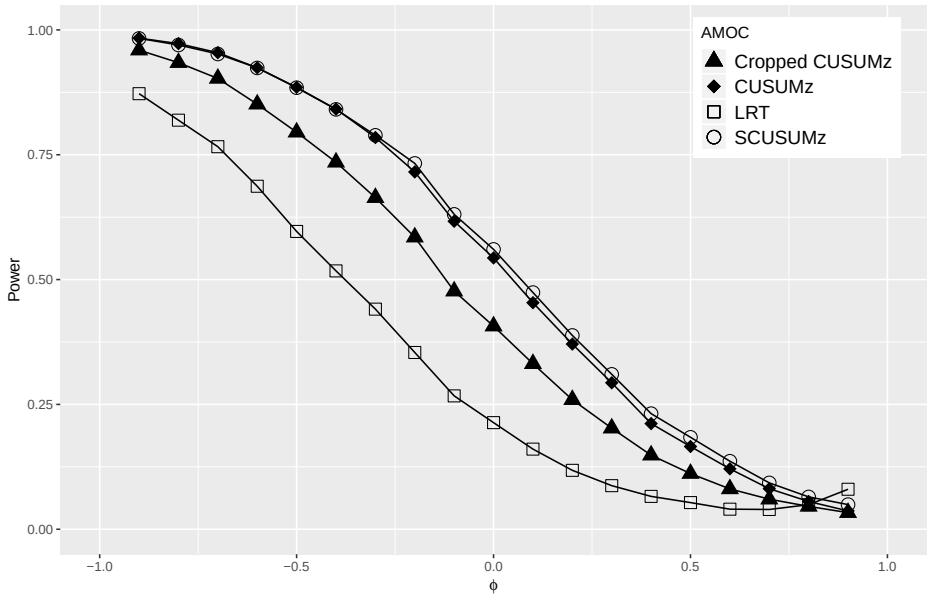


Figure 2.3: Detection Power for an AR(1) Series with Different ϕ. Here, $N = 1000$ and $\Delta = 0.15$.

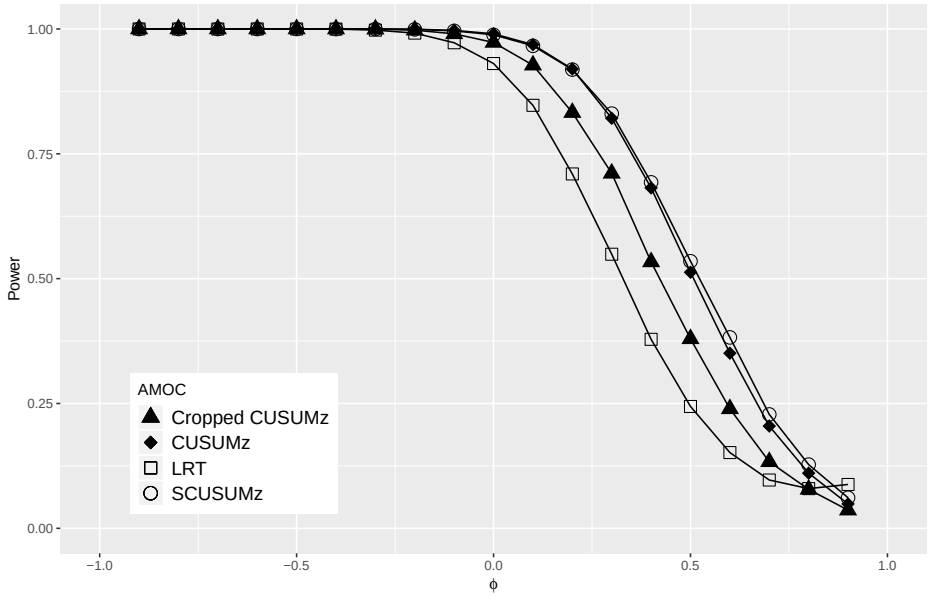


Figure 2.4: Detection Power for an AR(1) Series with Different ϕ. Here, $N = 1000$ and $\Delta = 0.3$.

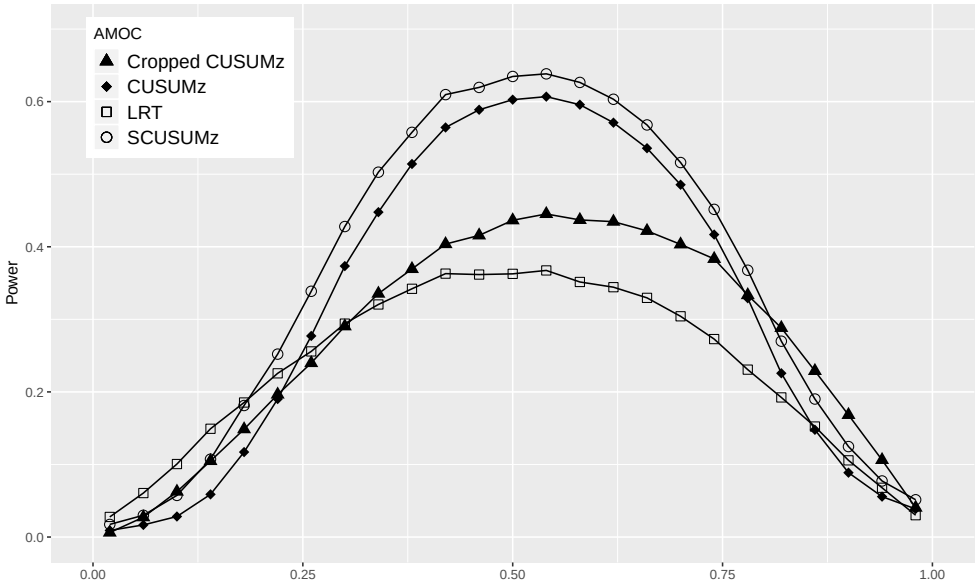


Figure 2.5: A Graph of $\frac{\tau}{N}$ Against Power with $N = 500$ and $\Delta = 0.5$ for an AR(1) Series with $\phi = 0.5$.

Table 2.3: Type I Error Estimation for AR(1) models, MA(1) models, ARMA$(1,1)$ Models, and ARMA$(2,2)$ Models. Here, $N = 1000$, $\sigma^2 = 1$, and $\alpha = 0.05$.

AR(1) Models

ϕ	CUSUMz	λ_z	SCUSUMz	LRT
-0.9	0.0451	0.0417	0.051	0.0103
-0.5	0.0449	0.0416	0.0498	0.0089
-0.25	0.0453	0.041	0.0509	0.0118
0.1	0.0431	0.0423	0.0449	0.0139
0.5	0.0388	0.038	0.0445	0.0156
0.7	0.0398	0.0377	0.0467	0.0274
0.9	0.0336	0.0298	0.0448	0.0749

MA(1) Models

θ	CUSUMz	λ_z	SCUSUMz	LRT
-0.95	0.0363	0.2257	0.0752	0.2733
-0.9	0.0393	0.0748	0.0567	0.1109
-0.5	0.0494	0.0466	0.0527	0.0123
0.1	0.0447	0.0435	0.0517	0.0123
0.5	0.0422	0.0438	0.0499	0.0171
0.9	0.0467	0.0439	0.0532	0.1133
0.95	0.0437	0.0433	0.0513	0.2469

ARMA$(1,1)$ Models

ϕ_1	θ_1	CUSUMz	λ_z	SCUSUMz	LRT
0.5	-0.95	0.0324	0.1698	0.0609	0.2952
0.5	-0.9	0.0373	0.06	0.0521	0.1232
0.5	-0.1	0.0426	0.0364	0.0482	0.019
0.1	-0.5	0.0424	0.0436	0.0476	0.0145
0.9	-0.5	0.0351	0.0333	0.0475	0.0607
0.95	-0.5	0.0328	0.0261	0.043	0.1281

ARMA$(2,2)$ Models

$\{\phi_1, \phi_2\}$	$\{\theta_1, \theta_2\}$	CUSUMz	λ_z	SCUSUMz	LRT
$\{0.6, 0.35\}$	$\{0.6, 0.3\}$	0.0309	0.0216	0.042	0.4874
$\{0.6, 0.3\}$	$\{0.5, -0.2\}$	0.0346	0.0272	0.0441	0.4344
$\{0.6, -0.1\}$	$\{-0.6, 0.3\}$	0.043	0.038	0.05	0.3473
$\{0.5, -0.2\}$	$\{-0.45, -0.5\}$	0.024	0.2065	0.0622	0.5768
$\{0.5, -0.2\}$	$\{-0.4, -0.5\}$	0.0371	0.0742	0.0591	0.4224
$\{0.2, -0.5\}$	$\{-0.45, -0.05\}$	0.0435	0.0424	0.0506	0.3433

Chapter 3

Multiple Changepoint Techniques

3.1 An Overview to the Multiple Changepoint Analysis

Now suppose that $\{X_t\}_{t=1}^N$ has an unknown number of changepoints, denoted by m, occurring at the unknown ordered times $1 < \tau_1 < \tau_2 < \cdots < \tau_m \leq N$. Boundary conditions take $\tau_0 = 1$ and $\tau_{m+1} = N + 1$. These m changepoints partition the series into $m + 1$ distinct regimes, the i^{th} regime having its own distinct mean and containing the points $\{X_{\tau_{i-1}}, \ldots, X_{\tau_i - 1}\}$. The model can be written as $X_t = \kappa_{r(t)} + \epsilon_t$, where $r(t)$ denotes the regime index, which takes values in $\{0, 1, \cdots, m\}$, and $\{\epsilon_t\}$ is a stationary causal and invertible ARMA(p, q) time series that applies to all regimes. Here, $\kappa_{r(t)} = \mu_i$ is constant for all times in the ith regime:

$$\kappa_{r(t)} = \begin{cases} \mu_0, & \tau_0 \leq t < \tau_1, \\ \mu_1, & \tau_1 \leq t < \tau_2, \\ \quad \vdots & \\ \mu_m, & \tau_m \leq t < \tau_{m+1}, \end{cases}$$

Multiple changepoint problems have received considerable attention over the past two decades. While the community has seemingly converged on single changepoint methods, there seems to be no concensus on how to tackle the multiple changepoint case. Two major camps have seemingly evolved.

First, the penalized likelihood camp minimizes a penalized likelihood objective function that

selects the number of changepoints and other model parameters. These methods are computationally intense as there are 2^{N-1} different admissible changepoint configurations to be explored in a sequence of length N (time one cannot be a changepoint). An exhaustive model search evaluating all possible changepoint configurations is not possible for even $N = 100$. Some authors [9, 23] use a genetic algorithm, which is an intelligent random walk search, to optimize the penalized likelihood. Unfortunately, such a randomized search may fail to identify a global minimum (converges to the optimal solution); moreover, one may need to wait several days for the genetic algorithm to converge when $T = 1000$. While penalized likelihoods can be used in a variety of scenarios, including time series errors, the camp argues about what penalty is optimal. Alternative approaches tackle the penalized likelihoods through the *dynamic programming*, for example, the **P**runed **E**xact **L**inear **T**ime (PELT) by Killick [20]. There is also a minority in the penalized likelihood camp which finds the changepoints by the ℓ_1 model selection, of which the penalties are post-tuning rather than pre-specified.

The second major camp takes a more algorithmic approach, attempting to devise estimation routines that run quickly and perform reasonably. The algorithmic approaches investigated by my dissertation are *binary segmentation* (BS) and *wild binary segmentation* (WBS). Binary segmentation, the earliest multiple changepoint algorithm, applies single changepoint techniques (AMOC) to identify the most prominent changepoint, and then splits the series into two subsegments about the flagged changepoint time (should it exist). The process is repeated iteratively to any subsegments until all subsegments are declared changepoint free. Binary segmentation is computationally fast and conceptually simple, but its nature as a greedy algorithm can leave it fooled. WBS [12] and narrowest-over-threshold detection [2] methods overcome binary segmentation weaknesses by drawing many small random subsegments, in hopes that a few of these subsegments will contain one and only one changpoint (this improves estimation). The injected randomness enables the binary segmentation to escape a local optimum.

While many authors have worked on the multiple changepoint issue, the audience should be aware that most of them assume i.i.d. $\{\epsilon_t\}$, for example, dynamic programming based approaches [20, 8], model selection using ℓ_1-regularization [40, 16], moving sum statistics [22] and binary/wild binary segmentation. Even in the AMOC case, it has been shown that techniques for independent data may not work well for time series with dependence [9, 23, 7].

As we will see in this dissertation, binary segmentation methods can be fooled; however,

they often give reasonable results and are almost trivial to compute. Penalized likelihood methods will in general give better results, but they are much more involved computationally.

While dealing with penalized likelihood approaches, a computational bottleneck arises. Because there are $\binom{N-1}{m}$ different admissible changepoint configurations in the series with m changepoints (time one cannot be a changepoint), there are 2^{N-1} different changepoint configurations to be considered when analyzing the entire series at once. This huge admissible model count can make an exhaustive model search — one that evaluates all admissible changepoint configurations — virtually impossible to conduct. Unfortunately, PELT by [20] and FPOP by [17], two dynamic programming based techniques, require the objective function to be additive over distinct regimes. Regime-additive likelihoods will not arise when $\{\epsilon_t\}$ is an ARMA(p, q) series. Perhaps more problematic, the time series parameters governing $\{\epsilon_t\}$ apply to all $m+1$ regimes and are estimated from all series values — one cannot rig up a dynamic programming scheme that bookkeeps a leftmost or rightmost changepoint time and "ignore" any data toward the boundaries. While we will use a genetic algorithm to optimize our penalized likelihoods, additional research is needed in optimization aspects of the penalized likelihoods.

3.2 Penalized Likelihood Methods

Penalized likelihood approaches analyze the whole series at once, optimizing a model likelihood function (e.g., maximizing a model likelihood) with a penalty term that controls the number of changepoints. The techniques seek a solution that minimizes the objective function

$$O(m; \tau_1, \ldots, \tau_m) = C(m; \tau_1, \ldots, \tau_m) + P(m; \tau_1, \ldots, \tau_m), \tag{3.1}$$

where C is the cost of a changepoint configuration and P is a penalty term to prevent over-tuning. There are many ways to define the cost and penalties. A frequently used cost is the negative likelihood or negative log-likelihood, which will be used here:

$$C(m; \tau_1, \ldots, \tau_m) = -2ln(L_{\text{opt}}(\theta|m; \tau_1, \ldots \tau_m)),$$

where $L_{\text{opt}}(\theta|m; \tau_1, \ldots, \tau_m)$ is the time series likelihood (Gaussian based) optimized over all parameters θ given that m changepoints occur at the times $\tau_1, \ldots, \tau_m$.

The penalty can be constructed in a variety of ways. Common penalties include minimum description lengths (MDL), modified Bayesian Information Criterion (mBIC), and the classic AIC and BIC penalties. Of these four penalties, the AIC and BIC penalties are simple multiples of the number of changepoints, while the MDL and mBIC further incorporate the changepoint times.

The MDL penalty is based on information theory and is discussed further in [35, 9] and [23]. It quantifies the minimum storage space that requires to store the parameters of a model. MDL essentially penalizes integer-valued parameters, such as the number and location of changepoints, more heavily than a real-valued parameter such as the variance of a series. Simply, it charges $\log(m)$ penalty for having m changepoints, $\sum_{i=1}^{m+1} \frac{1}{2}\log(\hat{\tau}_{i+1} - \hat{\tau}_i)$ penalty for estimating $m+1$ segment means, and $\sum_{i=1}^{m} \log(\hat{\tau}_i)$ penalty for m changepoint locations. Therefore,

$$\text{MDL}(\hat{\tau}_1, \cdots, \hat{\tau}_m) = \frac{N}{2}\ln(\hat{\sigma}^2) + \ln(m) + \frac{1}{2}\sum_{i=1}^{m+1}\ln(\hat{\tau}_i - \hat{\tau}_{i-1}) + \sum_{i=1}^{m}\ln(\hat{\tau}_i). \tag{3.2}$$

The mBIC penalty is developed in [49].

Except that AIC has notoriety for overestimating the number of model parameters, it is not clear which penalty will perform best. These penalties are listed in the following table for convenience.

Table 3.1: Penalized Likelihood Objective Functions

Criteria	Cost Function
AIC	$N\ln(\hat{\sigma}^2) + 2(2m+3)$
BIC	$N\ln(\hat{\sigma}^2) + (2m+2)\ln(N)$
mBIC	$\frac{N}{2}\ln(\hat{\sigma}^2) + \frac{3}{2}m\ln(N) + \frac{1}{2}\sum_{i=1}^{m+1}\ln\frac{(\hat{\tau}_i - \hat{\tau}_{i-1})}{N}$
MDL	$\frac{N}{2}\ln(\hat{\sigma}^2) + \ln(m) + \frac{1}{2}\sum_{i=1}^{m+1}\ln(\hat{\tau}_i - \hat{\tau}_{i-1}) + \sum_{i=1}^{m}\ln(\hat{\tau}_i)$

Here, $\hat{\sigma}^2$ is the estimated white noise variance of the innovations process in the ARMA(p,q) model fit.

Finding the changepoint configuration that minimizes the penalized likelihood is completed through a genetic algorithm search [9, 23]. A genetic algorithm is an intelligent random walk based search that is unlikely to evaluate changepoint configurations that are suboptimal. Because there are 2^{N-1} different admissible changepoint configurations to be considered, and evaluating each

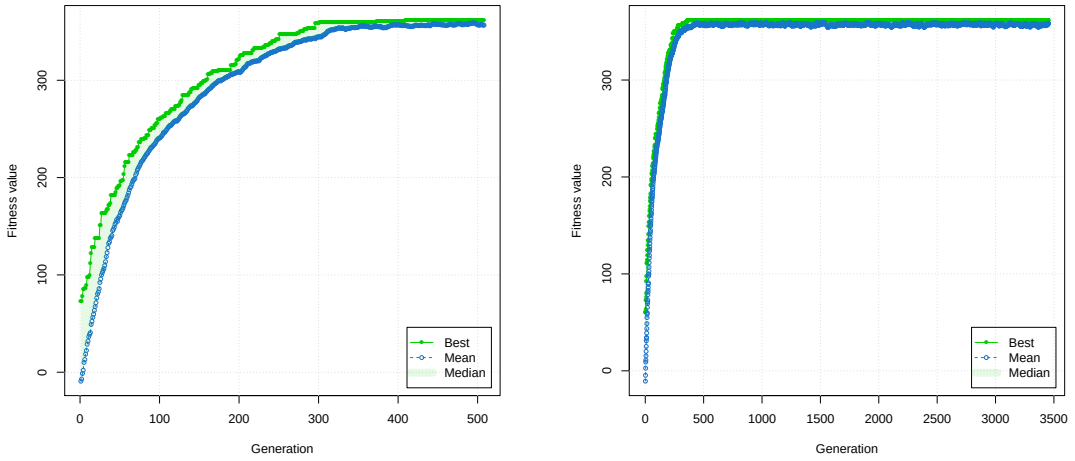
changepoint configuration entails optimizing a Gaussian time series likelihood, the optimization of a penalized likelihood is a non-trivial computation. Unfortunately, the objection function in (3.1) is not convex, and more research is needed in its optimization.

In this dissertation the changepoint time is encoded by a binary chromosome representation $\mathcal{G}$ of length N. The first bit of $\mathcal{G}$ is zero since it's never the changepoint time. For example, the following chromosome represents a sequence with two changepoints occurring at i^{th} and j^{th}:

$$\mathcal{G} = \{0, 0, \cdots, \underset{i}{1}, 0, \cdots, \underset{j}{1}, 0, \cdots, 0\}$$

In a genetic algorithm, an initial generation of candidate solutions (changepoint configurations) is randomly evolved toward better solutions; each candidate solution has a set of properties (like genes) that can be mutated and altered across each generation during the reproduction process; by simulating Darwin's theory of evolution and natural selection, only the optimal changepoint configuration will survive.

The choice of parameters for genetic algorithm is crucial. If the initial population size is too small, then the population will lose the diversity and the genetic algorithm is more likely to fall in local optimums; the bigger the population the better result; however if the population size is too big the search will be brute force. After considering every aspect and simulating for hundreds of different settings, the initial population size used in my dissertation is 200 given that the sequence length equals to 500. The mutation probability is $p = 0.05$. The maximum number of iteration decides the search depth and the number of generations tells when to stop the search if no improvements are found in consecutive generations. To guarantee the convergence of genetic algorithm in the changepoint analysis, these two parameters are chosen by intense scrutiny and are demonstrated by Figure 3.1



(a) Max of Iterations = 1000 and of Generations = 100 (b) Max Iterations = 30000 and Generations = 3000

Figure 3.1: Search Depth of Genetic Algorithm

Then the objective function is by `R GA` package from [39].

3.3 Binary Segmentation and Wild Binary Segmentation

Binary segmentation [38] estimates multiple changepoint configurations via AMOC methods. Binary segmentation is easy to understand: one first applies an AMOC technique to the entire series. When working with time series data, the AMOC technique should consider correlation. If a changepoint is declared, the algorithm splits the series into two subsegments about the flagged changepoint time. These two subsegments are subsequently analyzed by the AMOC technique for further changepoints. The process is iteratively continued until no subsegments are found to have a changepoint. Binary segmentation is simple to implement and computationally fast. However, it sometimes misses the global optimal solution as it is essentially a "greedy" algorithm that sequentially makes decisions based solely on the information at the current step. Binary segmentation also has difficulties handling small subsegments: AMOC conclusions are typically based on asymptotic results and the series' correlation structure may be inaccurately estimated from a short series segment.

Wild binary segmentation was developed in [12] and seeks to inherit the main computational

strengths of binary segmentation while eliminating its weaknesses of omitting possible changepoints. The technique works by drawing many segments $\{X_t\}_{s \le t \le e}$ such that (1) the start point s and end point e are selected independently with replacement uniformly from the times $\{2, \cdots, N\}$, and (2) $|e - s| \ge \delta_N$, where δ_N is a required prespecified minimum spacing between two changepoints. Next, an AMOC statistic for each subsample is computed, and the largest AMOC statistic over all segments is compared to the threshold

$$\tilde{\sigma} C \sqrt{2 \log(N)}, \tag{3.3}$$

where $\tilde{\sigma}$ is the median absolute deviation estimate of $\sqrt{\mathrm{Var}(\epsilon_t)}$, and C is a constant with a default value of 1 or 1.3. If this largest AMOC statistic exceeds the threshold, a changepoint is declared and the sequence is split into two subsegments about the estimated changepoint time. The same procedure is iteratively applied to any subsegments. The hope is that even a small number of random segments will contain a particularly 'favorable' segment in which the segment contains only one changepoint, sufficiently separated from both s and e. [12] suggests a lower bound for the number of draws (denoted by S) that guarantee such favorable draws with a high probability:

$$S \ge \frac{9N^2}{\delta_N^2} \log(N^2 \delta_N^{-1}). \tag{3.4}$$

The major difference between regular and wild binary segmentation is that regular binary segmentation applies a global CUSUM test to the whole series, while wild binary segmentation computes CUSUM for all randomly selected segments. Wild binary segmentation is in essence a randomized search, and the injected randomness enables wild binary segmentation to escape a local optimum and eventually achieve a global optimum. See [12] for additional detail. In our comparisons, the AMOC test adopted for binary and wild binary segmentation is the Sum of Squared CUSUM test, which essentially won the AMOC comparisons of the last two chapters.

3.4 A Further Discussion on WBS and WBS2

An improved version of the wild binary segmentation was published in early 2020 by Fryzlewicz [12] and we were invited to give a comment on WBS2 [26].

The WBS2 improvements made here are two-fold. The first simply selects the subintervals

to randomly sample in a data-driven way at stages, rather than generating all random subintervals at the algorithm's onset. When segments become small enough, all subsegments are sampled. This procedure yields an estimated changepoint configuration with $0, 1, \ldots, T-1$ changepoints. The second improvement replaced WBS thresholding methods to select the number of changepoints by a steepest-drop to low levels criteria that is shown to have some theoretical basis. The algorithm is dubbed WBS2-SDLL. For utility, many of our comments below apply both to WBS and WBS2-SDLL.

However, the pursuit of the high-frequency changepoint case seems somewhat over-emphasized. Indeed, with regard to the extreme teeth signal in Figure 1 of WBS2 article, a time series analyst would be remissed if their exploratory model/analysis neglected a seasonal (periodic) component. A power spectrum should readily identify the period of the data and a standard time series regression techniques would estimate the periodic mean and noise structure. One would obtain a more parsimonious model.

Of course, the extreme teeth signal could be replaced by one where the successive teeth had varying widths and heights to negate the above complaint, but in this case, other statistical techniques are available. Specifically, if one were given a time series with this structure, then a non-parametric regression analysis for the mean would be our first urge — especially in the absence of physical justification for the mean shifts. Phrased another way, an example where frequent mean shifts are physically plausible would be appreciated. The fitted changepoint configuration in the London House Price series in Figure 7 was not exciting to us: some of these changepoints seem more attributable to the positive correlations found in economic series than to true mean shifts.

For full disclosure, our interests in the multiple changepoint problem lie with climate time series, where weather stations have their gauges changed or are physically moved an average of six times per century in the United States [31]. This scenario dictates the need to allow for many abrupt changes in the mean. And because weather is correlated, one must allow for correlation in the model errors, something absent here (more on this below).

The above said, high dimensional series may well have many changes in its component series due to corporate mergers, political instability, or other equally vexing crises (imagine daily tracking of the 500 individual stocks in the S&P 500 stock index). As such, our concern is confined to the univariate setting.

Some of the claims that WBS methods work well in infrequent changepoint settings did not

jibe with our investigations of its performance. Elaborating, our most fundamental task typically involves the homogenization of a century of annually average temperatures at a station ($T = 100$). With this T, one typically has about three breaks for United States series (that is, roughly half of the gauge change and station move times induce a true mean shift), but sometimes there are none.

As such, a simulation was done on a time series with zero changepoint and simple i.i.d. $N(0, 1)$ noise with lengths $T = 100$ and $T = 500$. Here, WBS and WBS2-SDLL were compared to two often-used penalized likelihood methods: BIC and mBIC [49]. The results are summarized in the following table of average distances and empirical probabilities of getting one or more changepoints. These quantities were obtained over 1000 independent simulations.

Table 3.2: Average False Positive Rates and Distances

Methods	T=100		T=500	
	False Positive	Distance	False Positive	Distance
BIC	0.003	0.003	0.000	0.000
mBIC	0.034	0.040	0.006	0.006
WBS	0.204	0.390	0.079	0.103
WBS-SDLL	0.164	1.183	0.133	0.352

For a null hypothesis of no changepoints, a false positive rate of around 20 percent is not good. While WBS2-SDLL works better than WBS, performance issues remain. We refer to [42] for more comparisons of multiple changepoint techniques. Our overarching point is that a more extensive comparison of these techniques is needed in low- and mid-frequency changepoint settings. We would not care about performance in high-frequency settings if performance in low- and mid-frequency settings is sacrificed.

The above poor performance of WBS and WBS2-SDLL is likely still traced to threshold selection issues (or their equivalents). Towards this, additional rigorous probabilistic justification seems needed. Even in the original WBS setting, this does not appear to be an easy task. Developing this, a CUSUM changepoint statistic for Gaussian data computed over all admissible changepoint times in the interval (a, b) is distributed as the maximum absolute value drawn from a Gaussian process with a particular covariance structure, often expressed in limiting terms via the supremum of a Brownian bridge — see [37] for a recent treatment. The premise of WBS was to take many a and b uniformly (randomly) distributed over the observation times $\{1, 2, \ldots, T\}$. When one takes

a maximum over many (a, b), extreme value distributions would in principle would arise. But it is not clear how to proceed as correlation in the individual CUSUM maximums would exist should the corresponding intervals overlap; moreover, their individual distributions would depend on a and b, and there are many short intervals to deal with — the Brownian bridge approximation might be off.

As such, it is not clear how to proceed with WBS threshold selection on technical grounds. Likewise, the SDLL methods here sound interesting, but we had issues following the theoretical proofs. The claim that [46] has fixed all theoretical aspects of WBS was not appreciated here: Indeed, we got confused at ground zero with the asymptotic setup presented in both this and [46]. In particular, it would seem that as $T \to \infty$, the individual changepoint times η_i must depend on T for this scenario to make sense. But this is not stated; moreover, in such a scenario, we would need η_i to converge in some sense as $T \to \infty$. Consistently estimating any changepoint time η_i — in that its mean squared error goes to zero — should not be possible unless an infinity of observations is taken between all changepoint times. Related here: why would the mean shift sizes in Assumption 3.1 (d) need to depend on T? This seems unnecessary. Ditto assumption 3.1 (c) with a finite number of changepoints. We hope that the number of changepoints N is not changing in T? Infill asymptotics for multiple changepoint setup are considered in [9, 24]: the proofs are long and hard. In particular, we would appreciate discourse that illuminated all assumptions and steps showing where normality is needed, what maximal inequalities are needed, etc.

An assumption made here is that the sequence is i.i.d. and normally distributed. Independence is often questionable for time series data; moreover, neglecting correlation can severely influence multiple changepoint tests. As remarked above, the mean shifts in Figure 7 seem more attributable to correlation than mean shifts.

In this dissertation, if the variance parameter σ^2 were known, what would stop us from using PELT or related dynamic programming technique with a penalized likelihood for rapid computation? Indeed, it seems that a crucial component of the setup is how to estimate σ^2 accurately at the onset in the possible presence of many mean shifts. The authors mention MOSUM and some median absolute deviations of differences, but this seems to be key issue. In the case of time series data, how to accurately estimate an underlying autocovariance function in the series is going to be of paramount importance. Towards this, the methods here would breakdown if the autocovariance function changed at every changepoint time — as in the AR(1) segmentation literature [7].

As said in the introduction, the multiple changepoint literature is still in its infancy, and

methods that are computationally rapid and produce great segmentations across a robust variety of assumptions with appropriate technical justification have yet to be developed. The next decade will no doubt see much more work in this vein; this said, we believe that estimating the correct number of changepoints will be a tough task (as with most statistical smoothing problems).

The paper here has potential for us. WBS methods are known to be aggressive in that they tend to overestimate the number of changepoints. One could take the rapidly computed changepoint configuration here as a first step that could be further tuned with say some penalized likelihood method. Elaborating, if a series of length 10,000 could be reduced to say 100 good changepoint candidate times to explore, a genetic algorithm would make quick work of tuning up the configuration. Such an intitial configuration could also be used to place a prior on the changepoint times in the configuration as in [24].

3.5 A New Distance for Changepoint Technique Comparison

Several distances have been utilized by the multiple changepoint field. Some, such as the mean-squared-error (MSE) of the fitted means, V-measure, or Hausdorff distance are not specific to changepoints. Others, such as the number of changepoints or true/false positive rates of changepoint detection, are more tailored to the problem. However, each of the above distances typically quantifies only one aspect of the fit. For example, the MSE could be low, but the number of changepoints could still be overestimated; antipodally, the number of changepoints could be perfect, but their locations could be inaccurate. As such, we introduce a new changepoint-specific metric balancing the two key components of changepoint analysis: the number of changepoints and their locations. Note that the majority of changepoint approaches are not designed to optimize these aspects, but rather optimize a model fit or homogeneity of the parameter values. To balance the two critical aspects of numbers and locations of changepoints, two components in our distance are needed. The first is a measure of the discrepancy in the numbers of changepoints in the two configurations, for which we use absolute difference. The second component will measure the discrepancy in the location of the changepoints. This is tricker to quantify as the number of changepoints may be different in the two configurations and a "matching procedure" is needed.

Therefore, to better compare different changepoint techniques it will be useful to develop a distance between the two changepoint configurations $\mathcal{C}_1^m = (\tau_1, \ldots, \tau_m)$ and $\mathcal{C}_2^k = (\eta_1, \ldots, \eta_k)$,

where $\mathcal{C}_1^m$ and $\mathcal{C}_2^k$ are ordered sets. The distance between $\mathcal{C}_1^m$ and $\mathcal{C}_2^k$ is defined to be

$$d(\mathcal{C}_1^m, \mathcal{C}_2^k) = |m-k| + \min\{\mathcal{A}(\mathcal{C}_1^m, \mathcal{C}_2^k)\}. \tag{3.5}$$

The term $|m-k|$ assigns a score of one point for each discrepancy between the number of changepoints in the two configurations. The term $\min \mathcal{A}(\mathcal{C}_1^m, \mathcal{C}_2^k)$ reflects the minimum cost that matches changepoint locations from the set $\mathcal{C}_1^m$ with those from $\mathcal{C}_2^k$, thus can be computed via the following linear assignment

$$\begin{aligned} &\min \sum_{i=1}^{k}\sum_{j=1}^{m} c_{ij}x_{ij} \\ &\text{subject to} \sum_{i=1}^{k} x_{ij} = 1, \text{ for } j = 1, \cdots, m; \\ &\qquad \sum_{j=1}^{m} x_{ij} \leq 1, \text{ for } i = 1, \cdots, k; \\ &\qquad x_{ij} \in \{0, 1\}, \end{aligned} \tag{3.6}$$

where c_{ij} is the cost for assigning τ_i to η_j and is defined to be

$$c_{ij} = \frac{|\tau_i - \eta_j|}{N},$$

and x_{ij} is the decision variable such that

$$x_{ij} = \begin{cases} 1 & \text{if } \tau_i \text{ is assigned to } \eta_j \\ 0 & \text{otherwise} \end{cases}.$$

One can verify that the equation 3.5 defines a legitimate distance satisfying the triangle inequality. The term $\min \mathcal{A}(\mathcal{C}_1^m, \mathcal{C}_2^k)$ can be shown to be bounded by unity and measures how closely the two changepoint configurations with $\min(m, k)$ match up to one and other. The larger the distance is, the worse the two configurations correspond to one and other. Below, estimated multiple changepoint configurations will be compared to the true configuration with this distance. The linear assignment $\min \mathcal{A}(\mathcal{C}_1^m, \mathcal{C}_2^k)$ can be computed with efficient algorithms from [6]. Note that the solution may not be unique but the minimal cost is always unique, i.e., two different assignments give the

same minimal cost. Second, we are aware that the assignment problem has balanced ($m = k$) and unbalanced ($m \neq k$) cases. The unbalanced assignment problem is more complicated. If $m >> k$ or $k >> m$, the distance is simply approximated by $|m - k|$ which dominates the distance calculation; otherwise, $m \neq k$ results in a non-square cost matrix. To address the unbalanced case, $|m - k|$ "virtual nodes" with zero cost are added to the changepoint configuration with less changepoints so that a square cost matrix is obtained. For example, two changepoint configurations $\mathcal{C}_1^3$ and $\mathcal{C}_2^2$ for a time series of length 100 are $\mathcal{C}_1 = (25, 78, 99)$ and $\mathcal{C}_2 = (26, 51)$.

The cost matrix is

$$\text{cost matrix} = \begin{bmatrix} 1 & 26 \\ 51 & 27 \\ 73 & 48 \end{bmatrix} /100.$$

However, since it's unbalanced, we add a virtual "node"(changepoint) to $\mathcal{C}_2$ and the cost matrix becomes

$$\text{cost matrix} = \begin{bmatrix} 1 & 26 & 0 \\ 51 & 27 & 0 \\ 73 & 48 & 0 \end{bmatrix} /100.$$

With `R lpSolve` package the optimum solution to $\min \mathcal{A}(\mathcal{C}_1^3, \mathcal{C}_2^2)$ is found to be $\frac{28}{100}$ while the distance between $\mathcal{C}_1 = (25, 78, 99)$ and $\mathcal{C}_2 = (26, 51)$ is 1.28. If $\mathcal{C}_2 = (26, 51)$ represent the true changepoint times and $\mathcal{C}_1 = (25, 78, 99)$ is the result that identified by a changepoint technique, the distance 1.28 can be interpreted as the technique overestimates the number of changepoints by 1 which is the integer part of 1.28, and the distance of the estimated times to the true changepoint times is 0.28 which is the fraction part of 1.28.

3.6 Simulation Study on Multiple Changepoint Techniques

Our first simulation considers the changepoint free case in an AR(1) Gaussian series with various correlation parameters ϕ, $N = 500$, and $\sigma^2 = 1$. Figures 3.2 and 3.3 show probabilities of falsely declaring one or more changepoints and our distances averaged over 1,000 independent simulations.

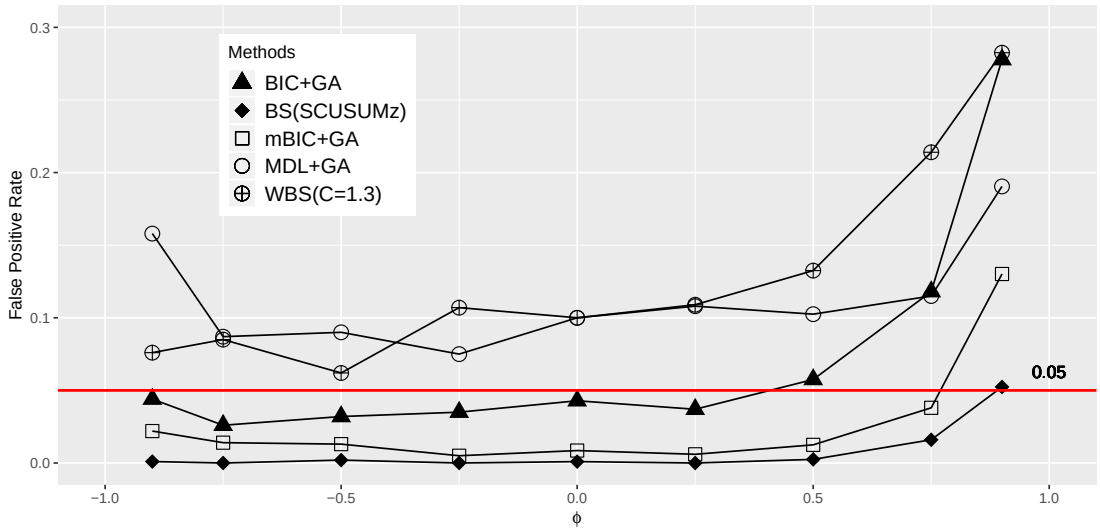


Figure 3.2: Empirical False Positive Detection Rates for an AR(1) Series with Various ϕ. Truth: No Changepoints.

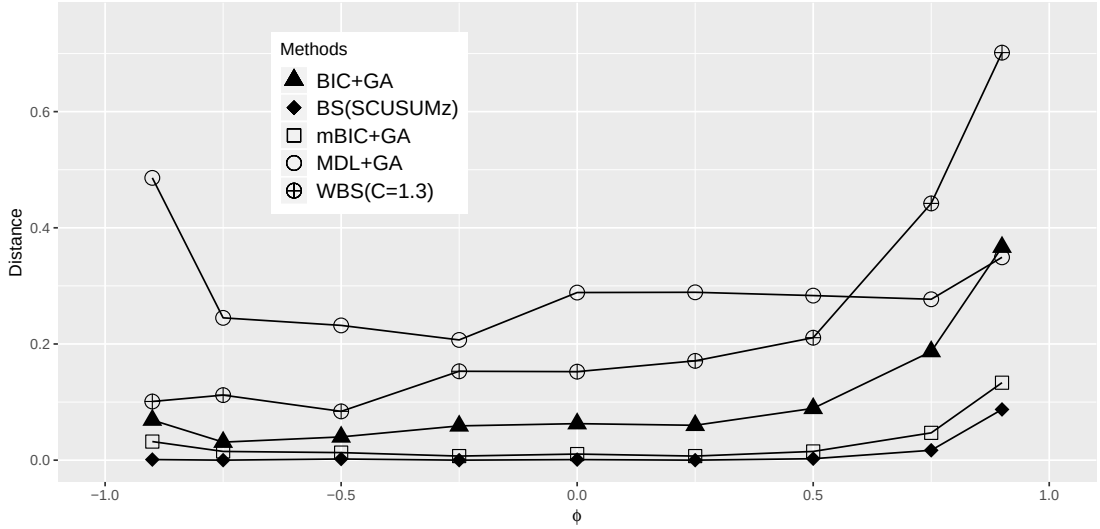


Figure 3.3: Average Distances for an AR(1) Series with Various ϕ. Truth: No Changepoints.

The results show that BIC, mBIC, and binary segmentation perform best, with WBS and MDL performing significantly worse. Binary segmentation performs best here, which is expected since there are no changepoints and a CUSUM type test applied to the whole series' one-step-ahead prediction residuals should not see a changepoint and stop any recursion from commencing. All methods perform better with negative ϕ than with positive ϕ; moreover, performance of all methods

degrades as ϕ moves upwards towards unity.

We now move to simulations with one changepoint in the same AR(1) setup above. The changepoint is put in the middle of the series (time 251) and has unit magnitude upwards. Figures 3.4, 3.5, and 3.6 show the proportion of runs estimating the correct single changepoint, the average number of changepoints estimated, and the distances between the estimated changepoint configurations and the true configuration.

In this case, WBS performance improves while MDL performance is still suboptimal. Binary segmentation, BIC, and mBIC all perform well across a large range of ϕ value; however, the performance of all tests again degrade as ϕ approaches unity.

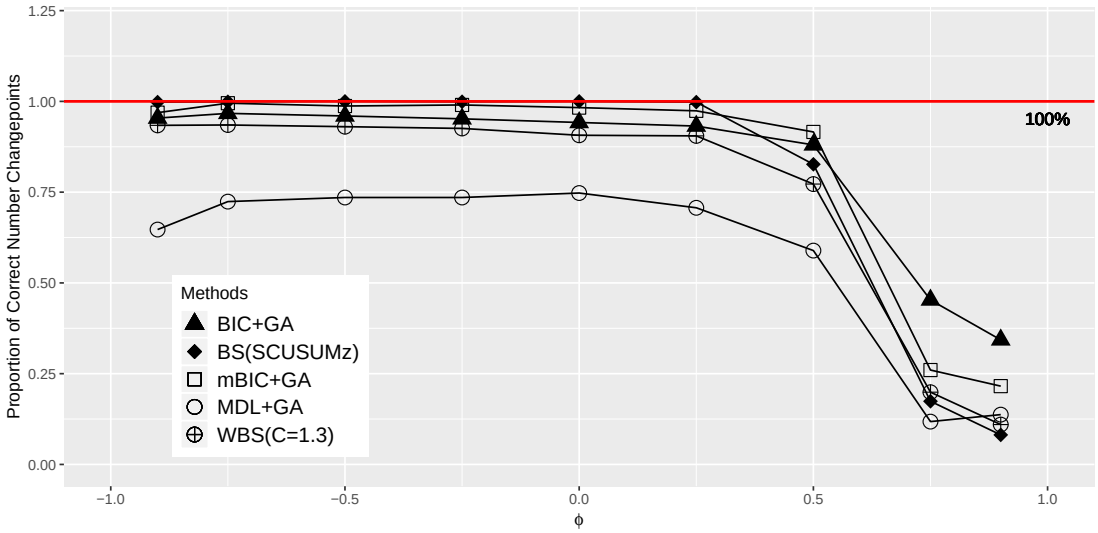


Figure 3.4: Proportion of Runs Correctly Estimating the Single Changepoint for an AR(1) Series with Varying ϕ. Truth: One Changepoint in the Middle Moving the Series Upwards.

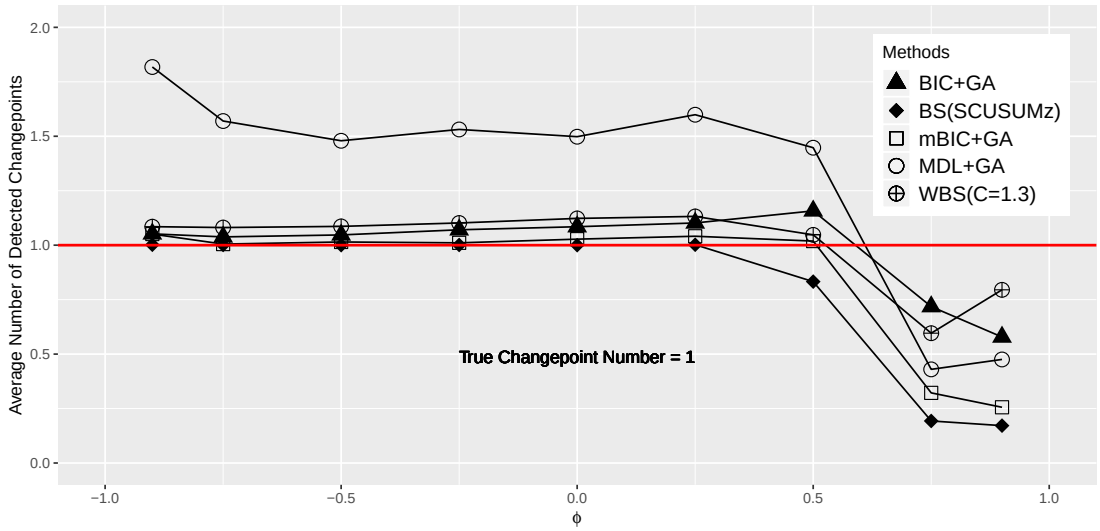


Figure 3.5: Average Number of Detected Changepoints for an AR(1) Series with Varying ϕ. Truth: One Changepoint in the Middle Moving the Series Upwards.

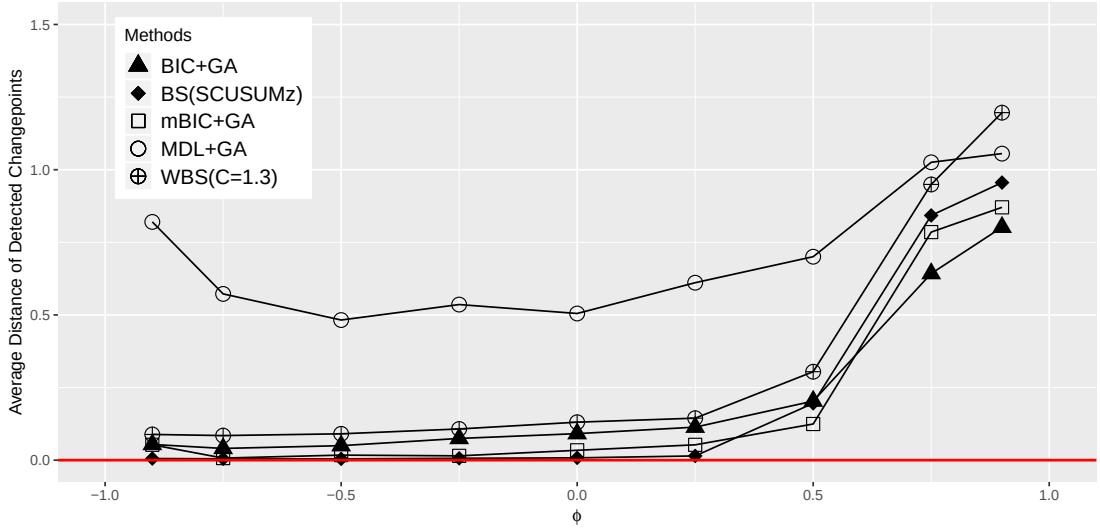


Figure 3.6: Average Distances for an AR(1) Series with Varying ϕ. Truth: One Changepoint in the Middle Moving the Series Upwards.

Our third simulation case moves to a setting with three mean shifts, partitioning the series into four equal-length regimes (the changepoints occur at times 126, 251, and 376), with each changepoint moving the series upward by one unit (up-up-up). Figures 3.10, 3.11, and 3.12 report results analogous to the single changepoint simulations. Many of the previous conclusions still hold. For example, all methods perform worse with positive ϕ than for negative ϕ. In this case,

WBS appears the worst and, except for MDL, the methods underestimate the correct number of changepoints.

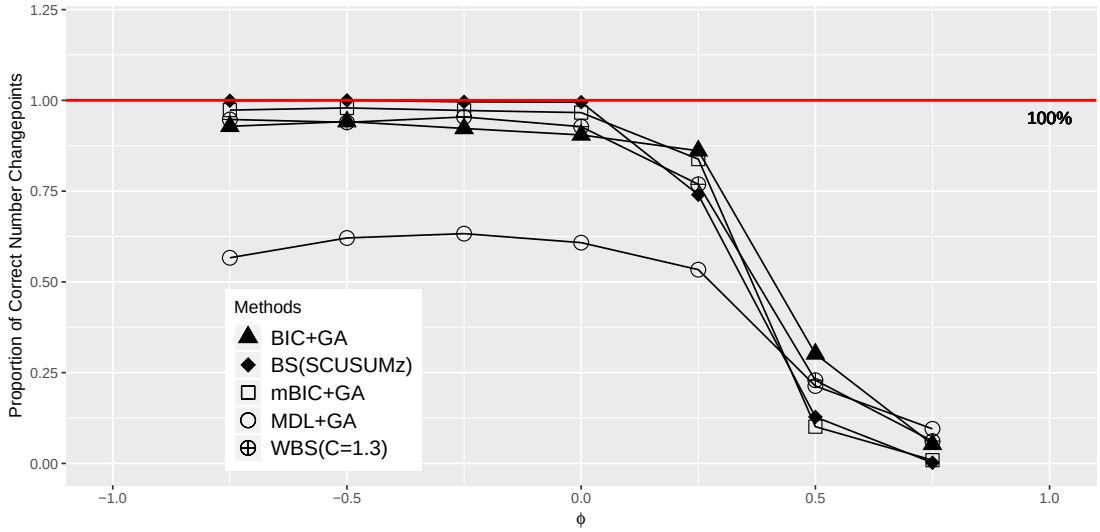


Figure 3.7: Proportion of Correctly Detecting the Changepoint Number for an AR(1) Series with Different ϕ. Truth: Three Equally Spaced Changepoints Moving the Series Up-Up-Up.

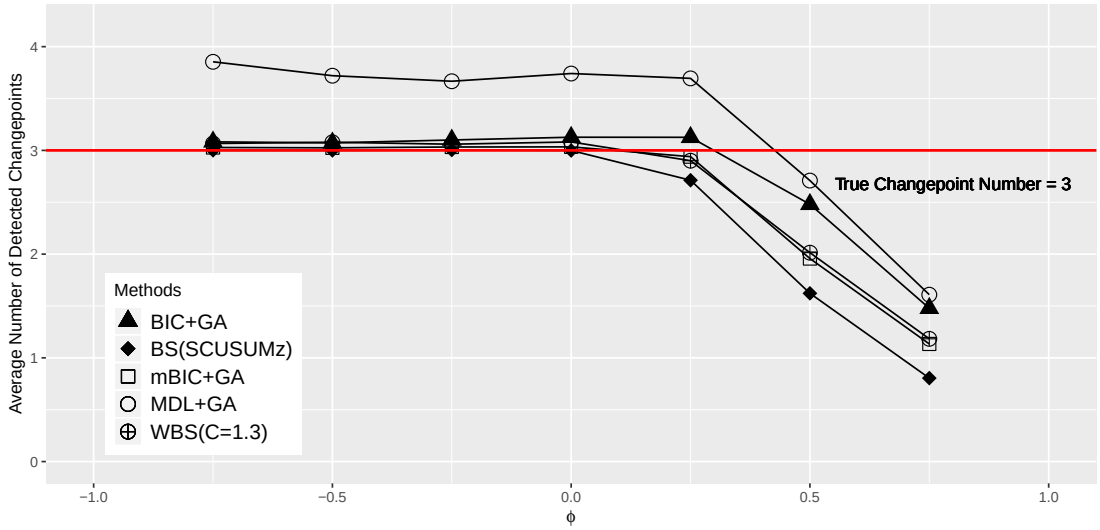


Figure 3.8: Average Number of Detected Changepoints for an AR(1) Series with Different ϕ. Truth: Three Equally Spaced Changepoints Moving the Series Up-Up-Up.

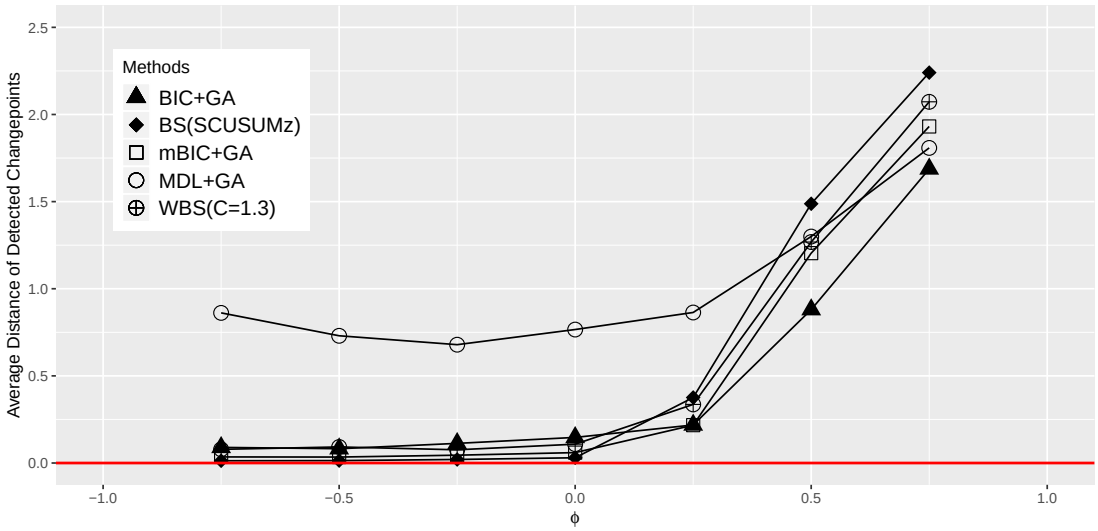


Figure 3.9: Average Distances for an AR(1) Series with Different ϕ. Truth: Three Equally Spaced Changepoints Moving the Series Up-Up-Up.

Next, we consider another three changepoint configuration, the changepoint times again being equally spaced with $N = 500$, but this time moving the series up, then down, and then up again (up-down-up). All mean shifts have a unit magnitude. Here, all methods have a harder time than the last Up-Up-Up three changepoint configuration. Tangible differences between the methods also become obvious. In this setting — as opposed to the up-up-up configuration above — binary segmentation becomes fooled and estimates too few changepoints. BIC performance begins to degrade as well. The better performing methods are BIC, mBIC, and binary segmentation.

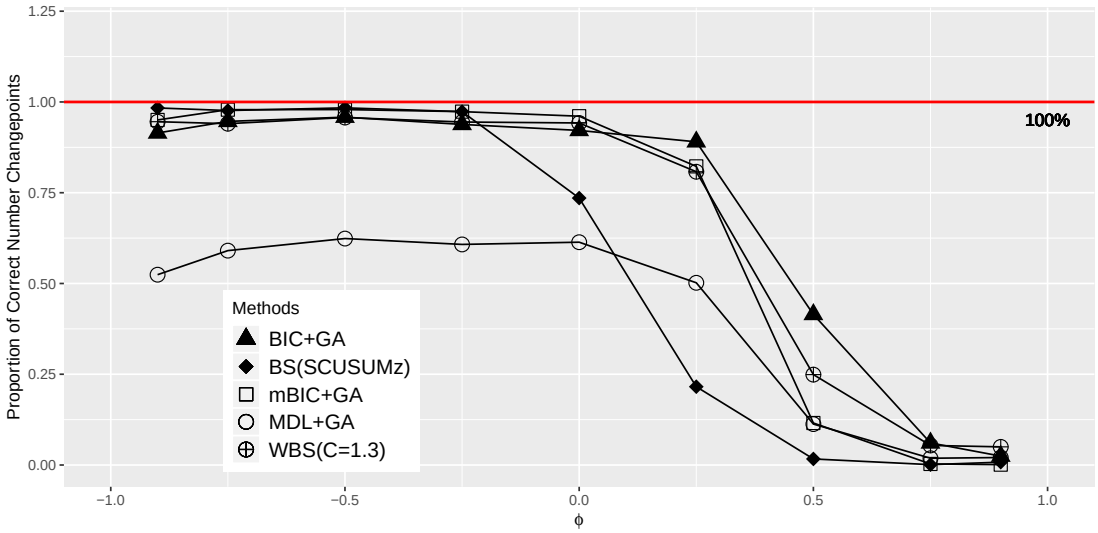


Figure 3.10: Proportion of Runs Correctly Estimating the Three Changepoints for an AR(1) Series with Varying ϕ. Truth: Three Equally Spaced Changepoints Moving the Series Up-Down-Up.

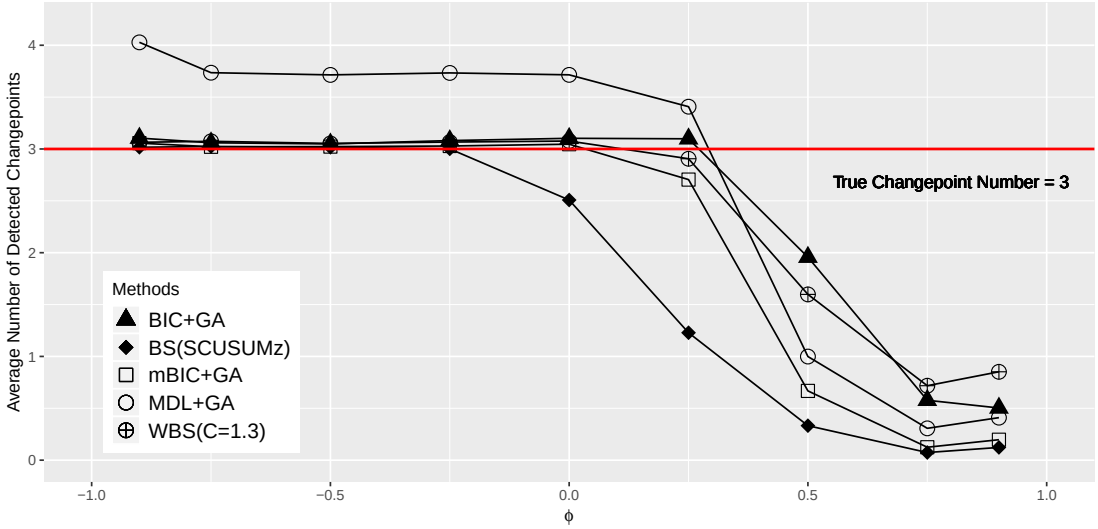


Figure 3.11: Average Number of Detected Changepoints for an AR(1) Series with Varying ϕ. Truth: Three Equally Spaced Changepoints Moving the Series Up-Down-Up.

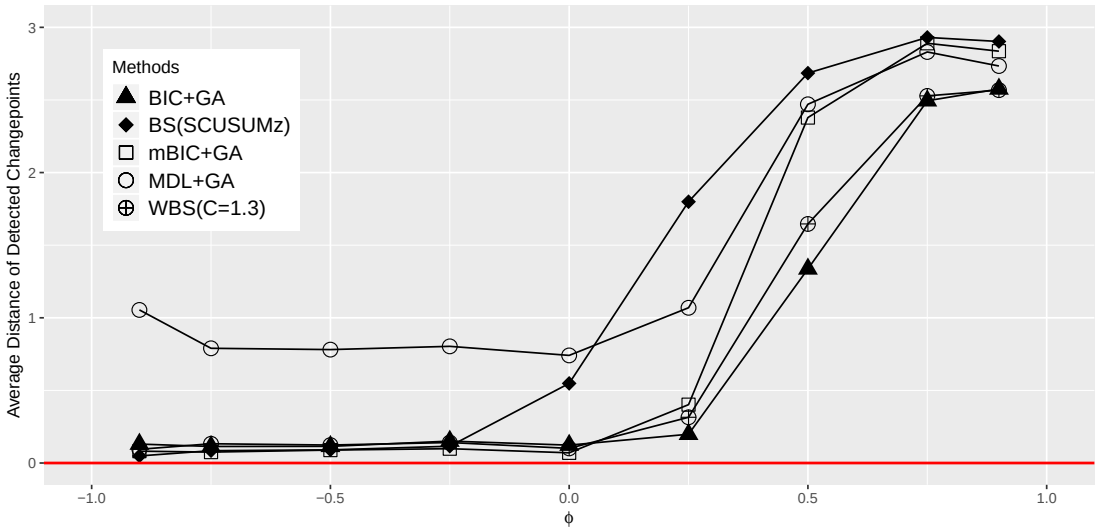


Figure 3.12: Average Distances for an AR(1) Series with Varying ϕ. Truth: Three Equally Spaced Changepoints Moving the Series Up-Down-Up.

At this point, it was surprising to us how well the simple BIC penalty has done — especially since this penalty dos not depend on the changepoint times. To examine this issue further, we fix the AR(1) parameter at $\phi = 0.5$ and study the behavior of the methods as N varies in the single changepoint case. Here, the changepoint is placed in the series' middle and moves it one unit higher. Table x reports averages distances of the methods when $N \in \{100, 500, 1000, 2500\}$. One sees that mBIC works better than BIC when the sequence is longer and the changepoint number is less. To our surprise that the classic BIC beats mBIC in all cases when the changepoint number is more. It seems that it does not help that mBIC incorporate the changepoint locations into the penalty. We have no idea what caused this but we will continue to investigate the reason. An initial guess is that mBIC was developed for the genomic data which are often long and have a small number of changepoints.

Table 3.3: Performance of multiple changepoint techniques when A changepoint is at middle and the sequence length N varies. $\sigma^2 = 1$, $\phi = 0.5$, $\Delta = 1$.

Results \ Methods	BIC+GA	mBIC+GA	MDL+GA	BS(SCUSUM)	WBS(C=1.3)
% Detect one cpt, $N = 100$	40.9%	30.3%	56.9%	21.9%	27.9%
($N = 500$)	(86.7%)	(91.3%)	(91.1%)	(81.3%)	(76.6%)
[$N = 1000$]	[93.2%]	[96.7%]	[90.7%]	[99.6%]	[91.9%]
{$N = 2500$}	{94.16% }	{ 97.72% }	{91.39% }	{100% }	{97.52% }
Avg # of detected cpts	1.436	0.530	3.291	0.252	0.738
	(1.176)	(0.986)	(1.167)	(0.817)	(1.021)
	[1.113]	[1.052]	[1.226]	[0.998]	[1.120]
	{1.115}	{1.042}	{1.209}	{1}	{1.027}
Avg distance to the true location	1.195	0.852	2.774	0.807	0.876
	(0.227)	(0.125)	(0.206)	(0.209)	(0.309)
	[0.126]	[0.066]	[0.239]	[0.017]	[0.135]
	{0.121}	{0.047}	{0.214}	{0.005}	{0.032}

Table 3.4: Performance of multiple changepoint techniques when three changepoints are equally spaced on the sequence with different lengths N. $\sigma^2 = 1$, $\phi = 0.5$, $\Delta's = 1$.

Results \ Methods	BIC+GA	mBIC+GA	MDL+GA	BS(SCUSUM)	WBS(C=1.3)
% Detect 3 cpts, $N = 100$	10.0 %	1.96 %	3.73 %	0.78 %	7.94%
($N = 500$)	(44.2%)	(10.1%)	(14.5%)	(1.8%)	(22.4%)
[$N = 1000$]	[82.8%]	[61.4%]	[56.7%]	[11.0%]	[67.0%]
{$N = 2500$}	{93.1% }	{97.1 % }	{85.7 % }	{76.3 % }	{98.2 % }
Avg # of detected cpts	1.25	0.40	3.03	0.13	0.85
	(2.04)	(0.63)	(1.31)	(0.31)	(1.54)
	[3.00]	[2.17]	[2.70]	[0.81]	[2.58]
	{3.10}	{3.03}	{3.35}	{2.55}	{3.02}
Avg distance to the true locations	2.57	2.90	3.98	2.88	2.35
	(1.27)	(2.42)	(1.95)	(2.70)	(1.69)
	[0.311]	[0.921]	[0.899]	[2.201]	[0.588]
	{0.123}	{0.066}	{0.364}	{0.486}	{0.036}

Table 3.5: Performance of multiple changepoint techniques when nine changepoints are equally spaced on the sequence with different lengths N. $\sigma^2 = 1$, $\phi = 0.5$, $\Delta's = 1$. $N = 100, 500$ are not simulated since more changepoints in a shorter series have a bigger impact on the autocorrelation estimate

Results \ Methods	BIC+GA	mBIC+GA	MDL+GA	BS(SCUSUM)	WBS(C=1.3)
% Detect 9 cpts, [$N = 1000$]	[1.49%]	[0.00%]	[0.00%]	[0.00 %]	[1.39 %]
{$N = 2500$}	{3.20% }	{9.41% }	{7.41% }	{2.50% }	{3.70% }
Avg # of detected cpts	[1.37]	[0.10]	[0.75]	[0.08]	[2.42]
	{8.06}	{3.45}	{5.55}	{0.51}	{6.99}
Avg distance to the true locations	[7.69]	[8.91]	[8.29]	[8.93]	[6.69]
	{1.29}	{5.66}	{3.96}	{8.50}	{2.31}

As perhaps our coup de grace scenario, we now move to cases with nine changepoints. Our first set of simulations equally spaces all changepoint times in the record with $N = 500$, each moving the series higher by one unit (All Up).

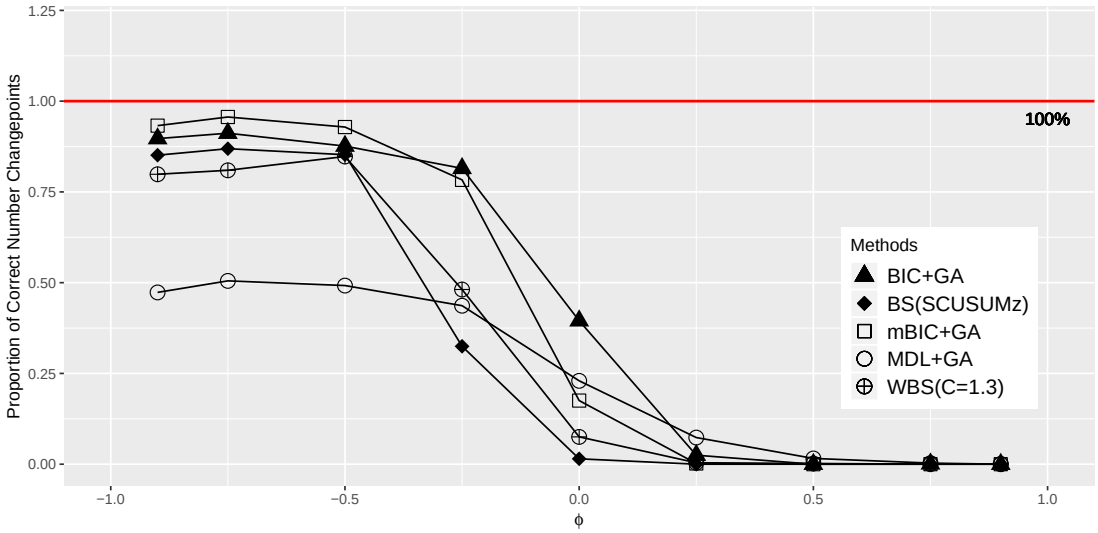


Figure 3.13: Proportion of Runs Detecting the Nine Changepoints for an AR(1) Series with Varying ϕ. Truth: Nine Changepoints, All Up.

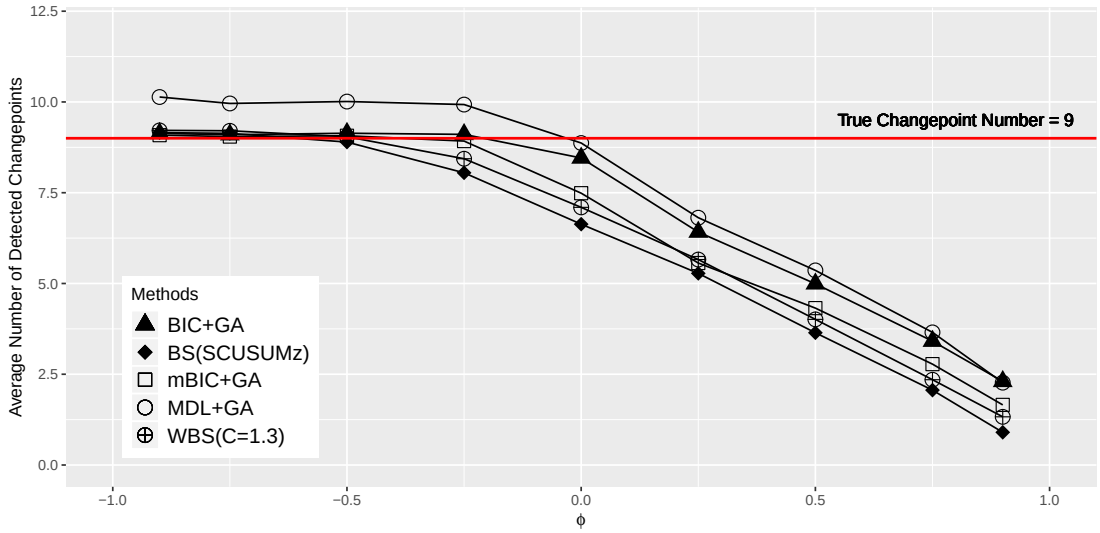


Figure 3.14: Average Number of Detected Changepoints for an AR(1) Series with Varying ϕ. Truth: Nine Changepoints, All Up.

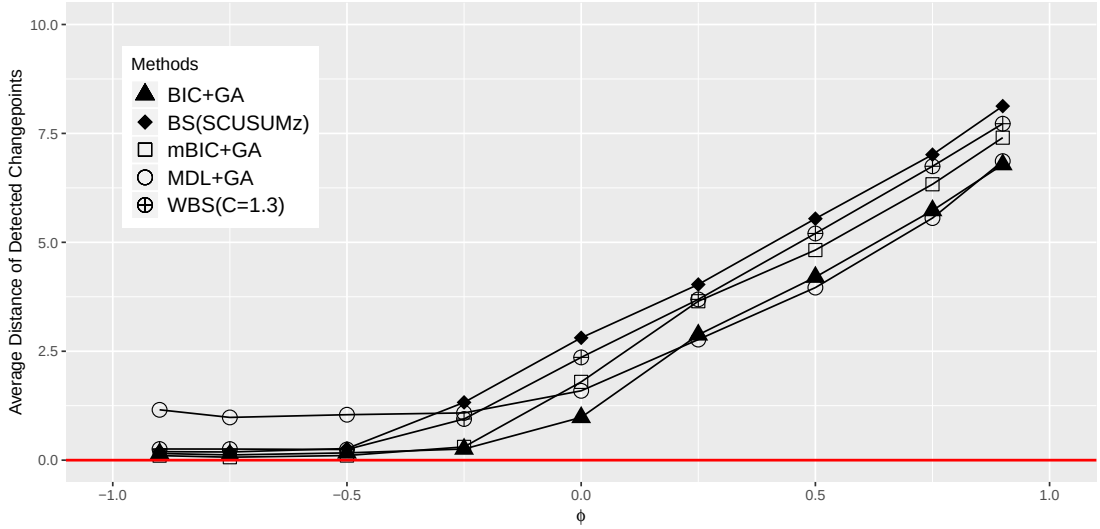


Figure 3.15: Average Distances for an AR(1) Series with Varying ϕ. Truth: Nine Changepoints, All Up.

Our next set of nine changepoints simply alternates the directions of the nine equally spaced unit mean shift sizes (Up-Down-Up-Down-Up-Down-Up-Down-Up), which we denote by Alternating.

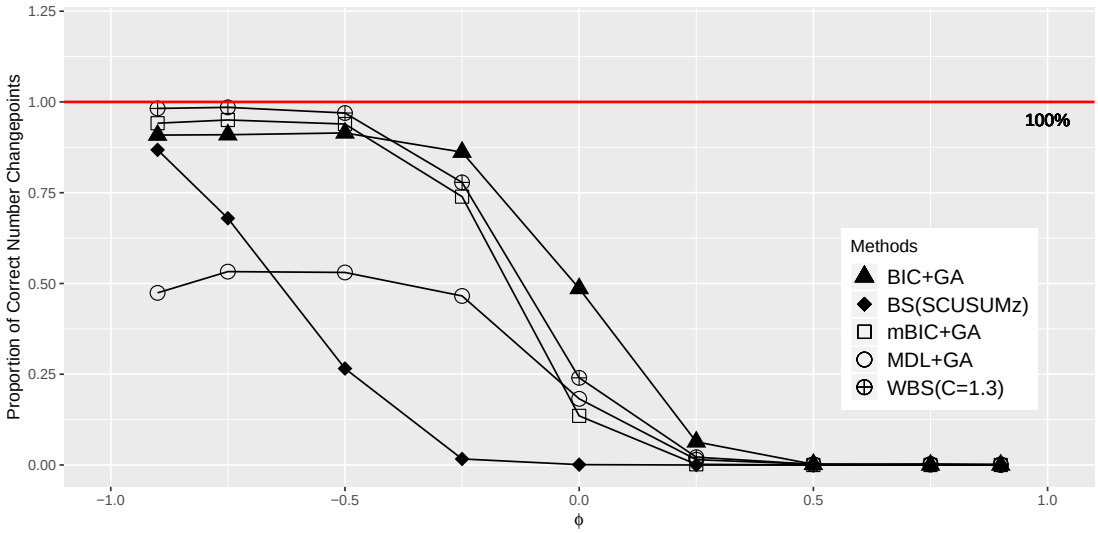


Figure 3.16: Proportion of Runs Correctly Detecting the Nine Changepoints for an AR(1) Series with Varying ϕ. Truth: Nine Alternating Changepoints.

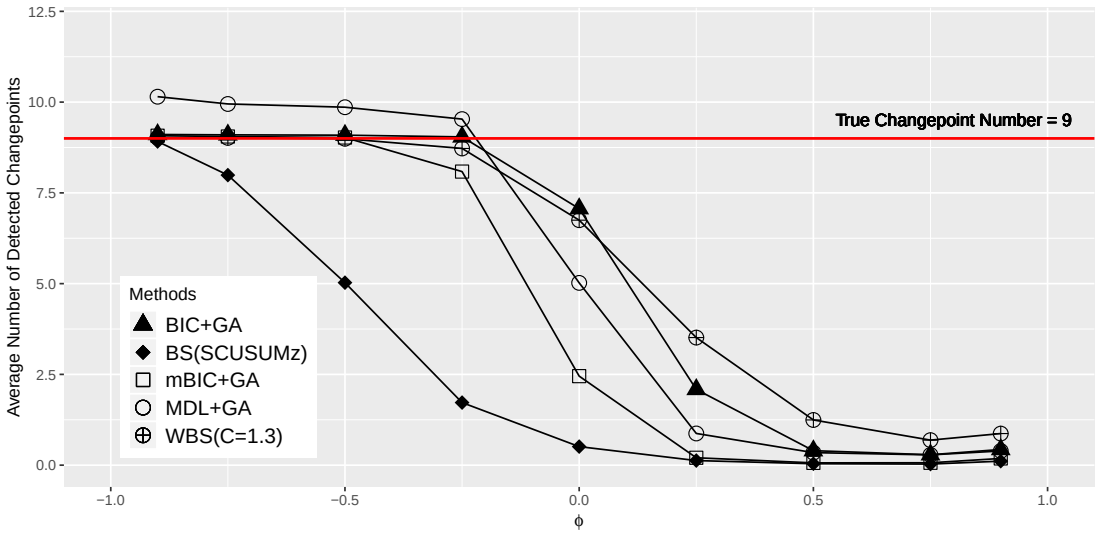


Figure 3.17: Average Number of Detected Changepoints for an AR(1) Series with Varying ϕ. Truth: Nine Alternating Changepoints.

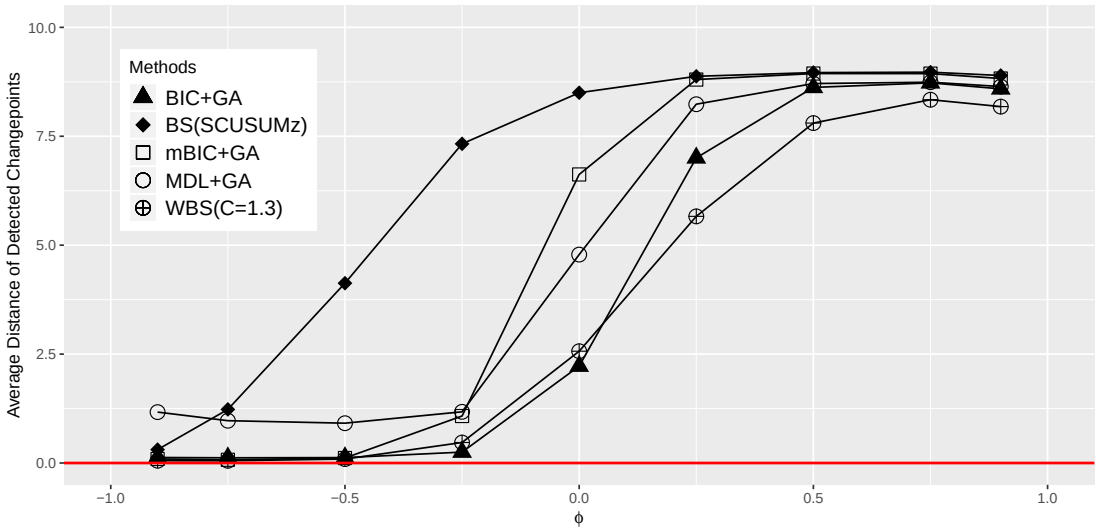


Figure 3.18: Average Distances for an AR(1) Series with Varying ϕ. Truth: Nine Alternating Changepoints.

The Keyblade setting contains nine changepoints with irregular locations and irregular mean shifts which are described by the figure 3.19:

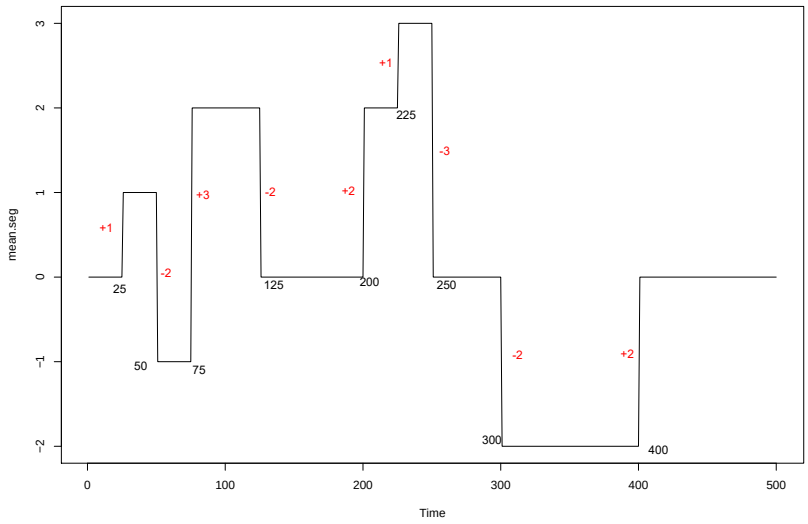


Figure 3.19: Changepoint Locations and Mean Shift Size of the Keyblade Signal. $\{\epsilon_t\}$ is an AR(1) Series with Varying ϕ.

Though MDL+GA or mBIC+GA performs slightly better than BIC at some values of ϕ from the statistics of proportion of runs correctly detecting nine changepoints and average number

of detected changepoints, BIC+GA has the smallest distance to the true changepoint locations, followed by MDL+GA.

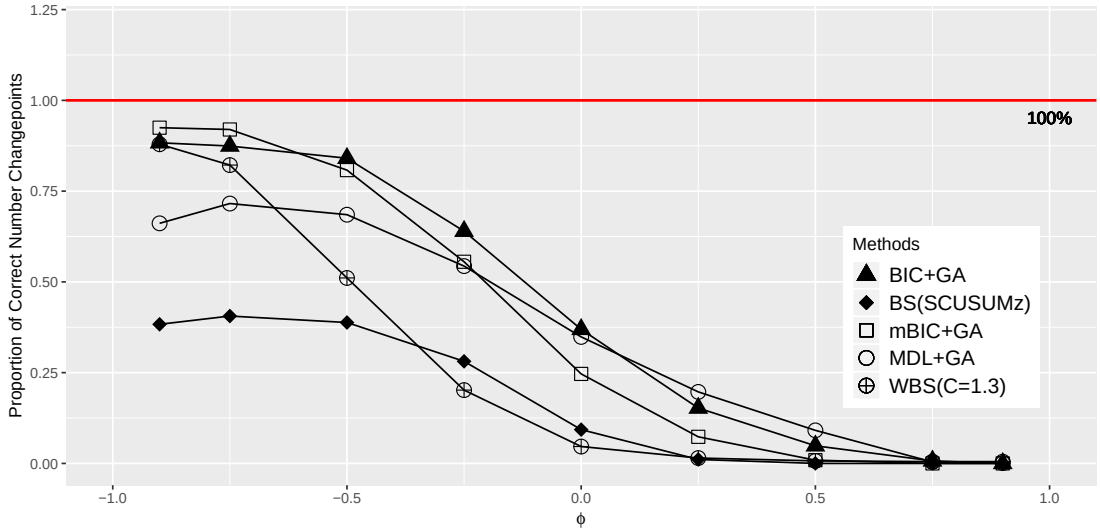


Figure 3.20: Proportion of Runs Correctly Detecting the Nine Changepoints for the Keyblade AR(1) Series with Varying ϕ. Truth: Nine Changepoints.

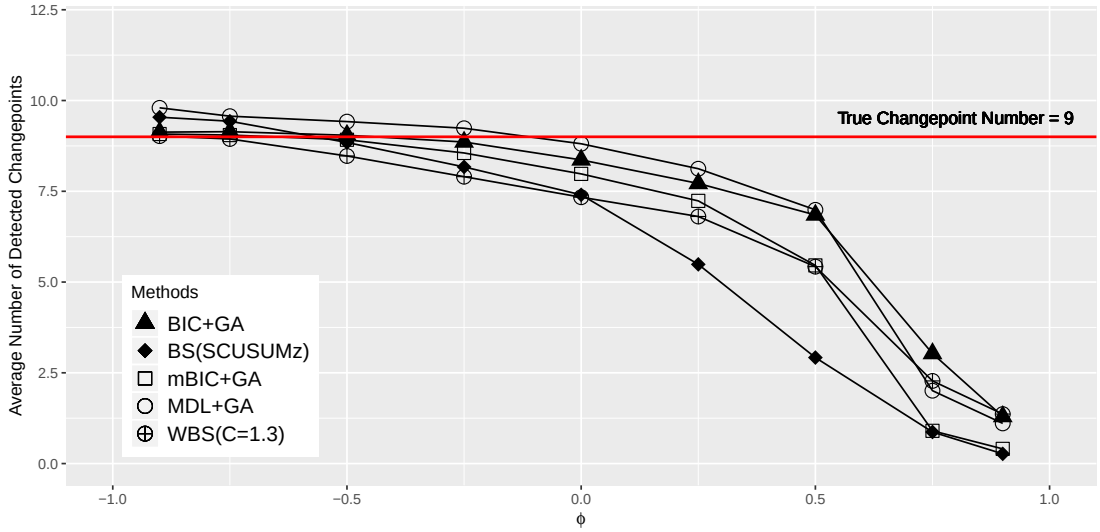


Figure 3.21: Average Number of Detected Changepoints for the Keyblade AR(1) Series with Varying ϕ. Truth: Nine Changepoints.

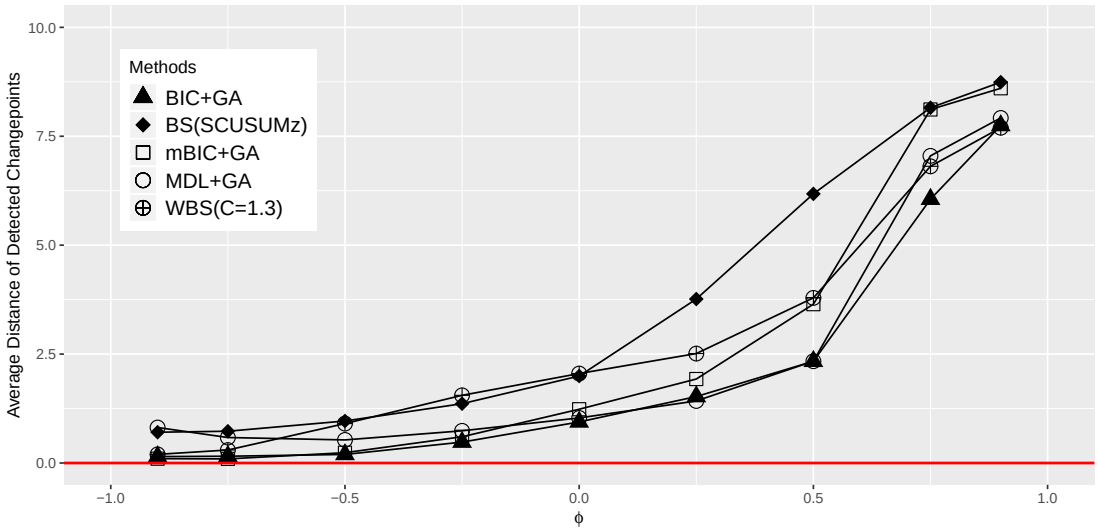


Figure 3.22: Average Distances for the Keyblade AR(1) Series with Varying ϕ. Truth: Nine Changepoints.

Chapter 4

Estimate the Autocovariance in Changepoint Problems

4.1 Autocovariance Estimation in Changepoint Problems

In section 3.4 we have stated the importance to accurately estimate an underlying autocovariance function of the time series in the changepoint analysis. A technique would signal false changepoints if the correlation structure was not removed from the series prior to the detection. The advantages of the autocovariance estimation in the changepoint analysis are not only to improve the accuracy of the detection but also to extend those i.i.d. based techniques to the correlated data, since the one-step-ahead residuals are asymptotically independent. This chapter describes a Yule-Walker type moment estimator to estimate the autocovariance structure for the changepoint problems. In addition to the inference, a simulation is conducted to show the Yule-Walker moment estimator works for the infrequent changepoint settings.

Reconsider the model defined in chapter 4:

$$X_t = \kappa_{r(t)} + \epsilon_t, \tag{4.1}$$

here $\{\epsilon_t\}$ is a stationary causal and invertible ARMA(p, q) time series that applies to all regimes

and obeys

$$\epsilon_t - \phi_1\epsilon_{t-1} - \cdots - \phi_p\epsilon_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \quad t \in \mathbb{Z}, \tag{4.2}$$

where $\{Z_t\}$ is IID with zero mean, variance σ^2, and a finite fourth moment. $r(t)$ denotes the regime index and takes values in $\{0, 1, \cdots, m\}$, and $\kappa_{r(t)} = \mu_i$ is constant for all times in the ith regime:

$$\kappa_{r(t)} = \begin{cases} \mu_0, & 1 \le t < \tau_1, \\ \mu_1, & \tau_1 \le t < \tau_2, \\ \quad \vdots & \\ \mu_m, & \tau_m \le t < N+1, \end{cases}$$

Chakar [7] took into account the AR(1) dependence structure in the changepoint detection and proposed a robust estimator of the AR(1) autocorrelation parameter as following:

$$\tilde{\rho}_n = \frac{\left(\underset{1\le t\le N-2}{\text{median}} |X_{t+2} - X_t|\right)^2}{\left(\underset{1\le t\le N-1}{\text{median}} |X_{t+1} - X_t|\right)^2} - 1. \tag{4.3}$$

Such an estimator has been proved consistent and satisfying the central limit theorem. However, it cannot be applied to autoregressive time series of higher orders ($p \ge 2$). Another changepoint literature dealing with the ARMA(p, q) time series are Davis and Lund [9]. The autocovariance estimate can be seen as a byproduct of the changepoint detection. However, it takes a significant amount of computational resource to obtain the estimates and thus it's not practical if the series is long. The third approach is to estimate the autocovariance function by a moving window of the length T such that $T < N$. The moving window generates $(N - T + 1)$ sub-segments, and each segment is treated as a stationary time series though some of them may contain mean shifts and thus cannot be stationary in fact. The parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$ of an ARMA(p, q) series are estimated $(N - T + 1)$ times and the medians are taken as the estimates of $(\boldsymbol{\phi}, \boldsymbol{\theta})$. However, the moving window approach still requires a full-scale investigation on the choice of the window length.

4.2 Yule-Walker Type Moment Estimators

To account for the impact of unknown mean shifts on autocovariance estimators, Yule-Walker type moment equations is proposed to estimate the autocovariance on the first order difference of $\{X_t\}$. The first order difference is calculated because the mean of $X_t - X_{t-1}$ is zero unless a changepoint occurs at time t. See Figure 4.1.

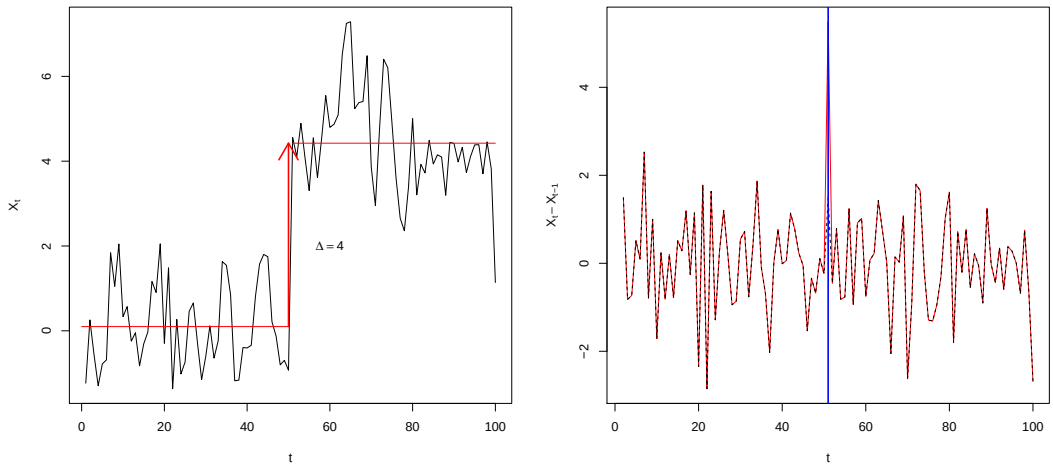


Figure 4.1: An AR(1) Series $\{X_t\}$ with a Changepoint at $t = 51$ (Left Panel) and its 1^{st} Order Differencing (Right Panel).

Define $d_t = X_t - X_{t-1} = \epsilon_t - \epsilon_{t-1}$, which holds except at time t that a changepoint occurs. Now we derive the moment equations for the coefficients of best linear predictors assuming an autoregressive structure of the order p. We use the following notation

$$\gamma(h) = E(\epsilon_t \epsilon_{t-h}),$$

$$d_t = \epsilon_t - \epsilon_{t-1},$$

$$\gamma_d(h) = Cov(d_t, d_{t-h}).$$

We can relate the autocovariance of a differenced series to that of the original series as follows:

$$\begin{aligned}\gamma_d(h) &= Cov(d_t, d_{t-h}) \\ &= Cov(Y_t - Y_{t-1}, Y_{t-h} - Y_{t-h-1}) \\ &= Cov(Y_t, Y_{t-h}) - Cov(Y_t, Y_{t-h-1}) - Cov(Y_{t-1}, Y_{t-h}) + Cov(Y_{t-1}, Y_{t-h-1}) \\ &= 2\gamma(h) - \gamma(h+1) - \gamma(h-1).\end{aligned}$$

Consider $r_k = \gamma(k) - \gamma(k-1)$ and

$$\begin{aligned}(i) \qquad & r_1 = -\frac{1}{2}\gamma_d(0) \\ (ii) \qquad & r_k = r_{k-1} - \gamma_d(k-1),\end{aligned}$$

Solve (i) and (ii), we have

$$r_k = \gamma(k) - \gamma(k-1) = -\frac{1}{2}\gamma_d(0) - \gamma_d(1) - \gamma_d(2) - \cdots - \gamma_d(k-1),$$

which implies

$$\frac{\gamma(k) - \gamma(k-1)}{\gamma_d(0)} = (-1)\left[\frac{1}{2} + \rho_d(1) + \cdots + \rho_d(k-1)\right]. \quad (\star)$$

If $\{\epsilon_t\}$ satisfies causal AR(p) difference equation (2.9) with $q = 0$, then the following hold

$$\begin{aligned}(a) \qquad & \gamma(h) = \phi_1\gamma(h-1) + \phi_2\gamma(h-2) + \cdots + \phi_p\gamma(h-p) && h = 1, \ldots, p \\ (b) \qquad & d_t = \phi_1 d_{t-1} + \cdots + \phi_p d_{t-p} + Z_t - Z_{t-1} && t = 0, \pm 1, \ldots \\ (c) \qquad & \gamma_d(h) = \phi_1\gamma_d(h-1) + \phi_2\gamma_d(h-2) + \cdots + \phi_p\gamma_d(h-p) && h = 2, \ldots, p,\end{aligned}$$

where (c) follows from (b). Subtracting equation (a) with $h = 2$ from equation (a) with $h = 1$ results in

$$\gamma(1) - \gamma(2) = \phi_1[\gamma(0) - \gamma(1)] + \phi_2[\gamma(1) - \gamma(0)] + \cdots + \phi_p[\gamma(p-1) - \gamma(p-2)].$$

Divide this equation by $\gamma_d(0)$ and use $(\star)$ to get our first moment equation relating the autocorrelation of the differenced series to the autoregressive coefficients.

$$\rho_d(1) + \frac{1}{2} = \frac{\phi_1}{2} - \frac{\phi_2}{2} - \phi_3\left[\frac{1}{2} + \rho_d(1)\right] - \cdots - \phi_p\left[\frac{1}{2} + \rho_d(1) + \cdots + \rho_d(p-2)\right],$$

where $\rho_d(h) = \frac{\gamma_d(h)}{\gamma_d(0}$. To get the remaining $p-1$ moment equations, we divide (c) by $\gamma_d(0)$. Our final system of linear equations becomes

$$\boldsymbol{\rho}_d = \boldsymbol{M}\boldsymbol{\phi}, \tag{4.4}$$

where

$$\boldsymbol{\rho}_d = \begin{bmatrix} \rho_d(1) + \frac{1}{2} \\ \rho_d(2) \\ \rho_d(3) \\ \vdots \\ \rho_d(p) \end{bmatrix} \qquad \boldsymbol{\phi} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \vdots \\ \phi_p \end{bmatrix},$$

and

$$\boldsymbol{M} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & -\left(\frac{1}{2} + \rho_d(1)\right) & \cdots & -\left(\frac{1}{2} + \sum_{j=1}^{p-2}\rho_d(j)\right) \\ \rho_d(1) & \rho_d(0) & \rho_d(1) & \cdots & \rho_d(p-2) \\ \rho_d(2) & \rho_d(1) & \rho_d(0) & \cdots & \rho_d(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_d(p-1) & \rho_d(p-2) & \rho_d(p-3) & \cdots & \rho_d(0) \end{bmatrix}$$

To estimate the autoregressive parameters in practice, we difference observed series $X_1, \ldots, X_n$, and calculate the sample autocorrelation of these differences up to lag p, our estimator becomes

$$\hat{\boldsymbol{\phi}} = \hat{\boldsymbol{M}}^{-1}\hat{\boldsymbol{\rho}}_d, \tag{4.5}$$

where in the elements of $\hat{\boldsymbol{M}}$ and $\hat{\boldsymbol{\rho}}_d$ we replace $\rho_d(h)$ with estimator

$$\hat{\rho}_d = \frac{\hat{\gamma}_d(h)}{\hat{\gamma}_d(0)} = \frac{\sum_{t=2}^{n-h}(X_t - X_{t-1})(X_{t+h} - X_{t+h-1})}{\sum_{t=2}^{n}(X_t - X_{t-1})^2}.$$

If the number of changepoints m is small relative to the sample size n, then the mean shifts will have negligible impact on the estimated covariance of the differences, since $X_t - X_{t-1} = d_t - d_{t-1}$ except at the changepoint times $\tau_1, \ldots, \tau_m$.

We end this section with a discussion of as estimate of white noise variance. Multiplying both sides of (b) above by d_{t-1}, taking expectations and solving for the $\text{var}(Z_t) = \sigma^2$, yields

$$\sigma^2 = \left(\sum_{j=1}^{p} \phi_j \gamma_d(j-1) \right) - \gamma_d(0).$$

A moment based estimator of the variance is given by

$$\hat{\sigma}^2 = \left(\sum_{j=1}^{p} \hat{\phi}_j \hat{\gamma}_d(j-1) \right) - \hat{\gamma}_d(0). \tag{4.6}$$

Based on the results in the next section, this above is a consistent estimator of the white noise variance.

4.3 Asymptotic Normality

In this section we show that if the number of changepoints $m = m(n)$ grows with the sample size slowly, that the parameter estimates derived in the previous section will be consistent and asymptotically normally distributed. We begin by describing the asymptotic normality of the autocorrelation for first differences in the general ARMA(p, q) case, which may be of independent interest. The asymptotic normality of the autoregressive parameter moment estimators follows as a corollary to Theorem 5.

Theorem 5. *If $\{X_t\}$ follows model (4.1) with $\{\epsilon_t\}$ satisfying (2.9), then for each positive integer k as $n \to \infty$,*

$$\sqrt{n}\left(\hat{\rho}_d(1), \ldots, \hat{\rho}_d(k) - (\rho_d(1), \ldots, \rho_d(k)\right)^T \Rightarrow N_m(\mathbf{0}, \mathbf{AWA^T}),$$

here the elements of the $(k+1)\times(k+1)$ matrix W are given by Bartlett's formula, e.g., see Brockwell

and Davis (1998), Chapter 8, while the matrix **A** *has form*

$$\boldsymbol{A} = \frac{1}{2(1-\rho(1))} \begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 \end{bmatrix}$$

Proof First we show that the changepoints have negligible impact on the estimated autocorrelations in the limit. Toward this end, let

$$\tilde{d}_t = X_t - X_{t-1} = (\epsilon_t - \epsilon_{t-1}) + \delta_t,$$

with

$$\delta_t = (\mu_k - \mu_{k-1})I_t = \tau_{k+1}.$$

As in the previous section we denote the calculated sample autocovarince of the observed differences as

$$\hat{\gamma}_d(h) = \frac{\sum_{t=2}^{n-h}(X_t - X_{t-1})(X_{t+h} - X_{t+h-1})}{n}.$$

If we let

$$\tilde{\gamma}_d(h) = \frac{\sum_{t=2}^{n-h}(\epsilon_t - \epsilon_{t-1})(\epsilon_{t+h} - \epsilon_{t+h-1})}{n},$$

then

$$\sqrt{n}|\hat{\gamma}_d(h) - \tilde{\gamma}_d(h)| \quad \leq \quad \frac{m}{\sqrt{n}} \left[m^{-1} B \sum_{t=\tau_j} (2|d_t| + B) \right],$$

where $B = \max_{0 \leq k \leq m} |\mu_k - \mu_{k-1}|$ is the maximum mean shift. The term on the right hand side converges to zero in the almost sure sense if $n^{-1/2}m \to 0$, as $n \to \infty$. Now we find the asymptotic distribution of the autocorrelation of the differences of $\{\epsilon_t\}$. Note that

$$\rho_d(h) = \frac{-\rho(h-1) + 2\rho(h) - \rho(h+1)}{2 * (1 - \rho(1))}, \qquad \hat{\rho}_d(h) = \frac{-\hat{\rho}(h-1) + 2\hat{\rho}(h) - \hat{\rho}(h+1)}{2 * (1 - \hat{\rho}(1))}.$$

The result of Theorem 5 now follows by the mapping theorem and well known results for asymptotic normality for sample autocovariances for ARMA processes, as described in chapter 8 of [5].

Corollary 6. *Let $\{X_t\}$ follows model (4.1) with $\{\epsilon_t\}$ satisfying (2.9) with $q = 0$. For estimator given in (4.5), as $n \to \infty$,*

$$\sqrt{n}\left(\hat{\phi}_1, \ldots, \hat{\phi}_p) - (\phi_1, \ldots, \phi_p\right)^T \Rightarrow N_m(\mathbf{0}, \mathbf{\Sigma}),$$

where T is the transpose operator and $\Rightarrow$ denotes convergence in distribution. Here

$$\mathbf{\Sigma} = \boldsymbol{MAW}(\boldsymbol{MA})^T.$$

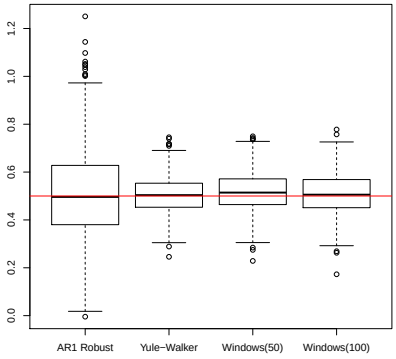
Proof of the Corollary 6 . Since the series of differences, $\{d_t = \epsilon_t - \epsilon_{t-1}\}$, is stationary and ergodic, the elements of $\hat{\boldsymbol{M}}$ converge to those of $\boldsymbol{M}$ in the almost sure sense. The conclusion of Corollary 6 follows by another application of the mapping theorem.

4.4 A Simulation Study

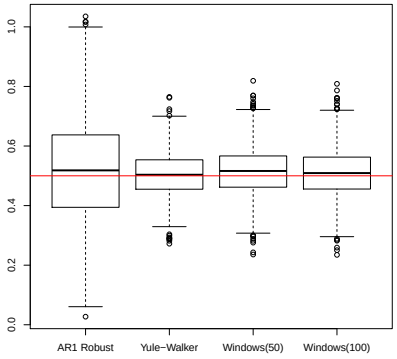
A simulation study is designed to examine the performance of the Yule-Walker estimator. All simulations are conducted on 1000 randomly generated time series of length $N = 500$, and $X_t = \mu_{r(t)} + \epsilon_t$ where ϵ_t follows an AR(p) process with $Var(Z_t) = 1$. The first case simulated is $\epsilon_t \sim$ AR(1) with $\phi = 0.5$. In this cases four competitive methods, AR(1) Robust estimator, Yule-Walker, rolling window of length 50 and 100, are examined on four changepoint settings: zero changepoint, one changepoint at the middle of series, four changepoints equally spaced on the series with a mean shift size of 2 and twenty five changepoints equally spaced on the series with a mean shift size of 1. The red horizontal line in each boxplot represent the true value of ϕ, which is 0.5. It's clear that the Yule-Walker estimator outperforms other estimators in various settings. As the mean shift size or the changepoint number increases, the bias of all estimators increases.

Next we turn to simulate on the AR(4) time series with true parameters $\phi_1 = 0.5$, $\phi_2 = -0.4$, $\phi_3 = 0.6$ and $\phi_4 = -0.3$. At present there are no competitive estimators for higher order autoregressive time series so the estimates are compared to the true parameters which are highlighted by red horizontal lines in each boxplot. It can be seen that the Yule-Walker moment estimator works

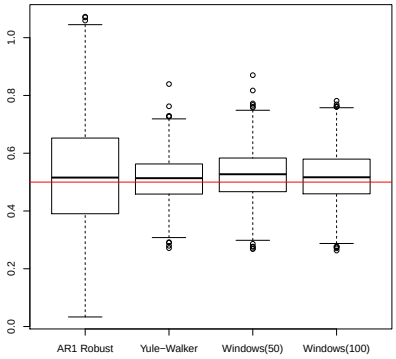
well when the changepoints are infrequent.



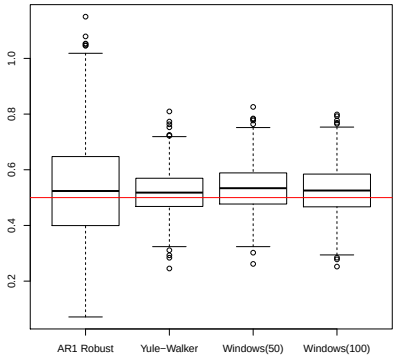
(a) No Changepoint



(b) One changepoint at $t = 251$, $\Delta = 2$

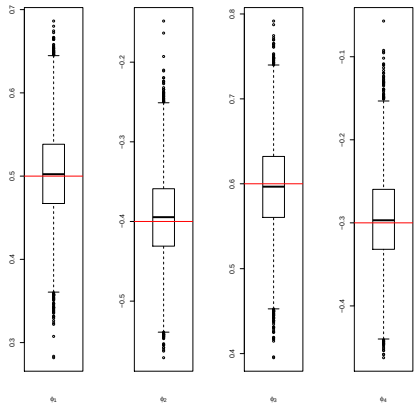


(c) Four changepoints occur at $t = 101, 201, 301, 401$ with equal shift sizes $\Delta = 2$

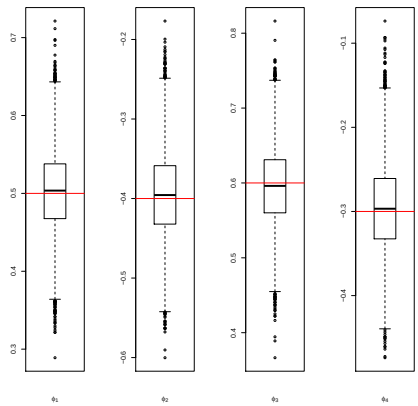


(d) Twenty five changepoints occur at $t = 21, 41, \cdots, 461, 481$ with equal shift sizes $\Delta = 1$

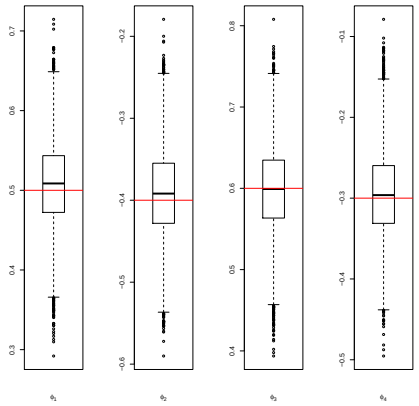
Figure 4.2: Autocorrelation Estimates for an AR(1) Series using Different Estimators.



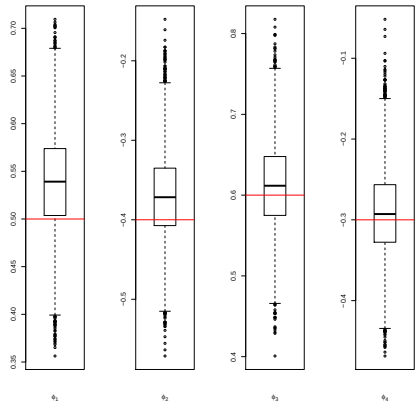
(a) No Changepoint



(b) One changepoint at $t = 251$, $\Delta = 1$



(c) Four changepoints occur at $t = 101, 201, 301, 401$ with equal shift sizes $\Delta = 1$



(d) Twenty five changepoints occur at $t = 21, 41, \cdots, 461, 481$ with equal shift sizes $\Delta = 1$

Figure 4.3: Autocorrelation Estimates for an AR(4) Series using Yule-Walker Estimator.

Chapter 5

Changepoint Detection using ℓ_1-Regularization

The attempt to apply the model selection approaches to the changepoint detection was made one decade ago. The underlying idea is that we can treat a possible changepoint as a feature of the model, then model selections, for example, ℓ_1-regularization, can be deployed to decide whether to include a feature into the model(a changepoint) or reject it. Research has been done both in theory and application, for example, adaptive LASSO [40] and total variation [16]. This chapter reviews the application of ℓ_1-regularization in the changepoint analysis.

ℓ_1-regularization methods are distinct from above-mentioned changepoint techniques. ℓ_1 approaches are simultaneous which estimate the size and time of mean shifts at the same time. In contrast, other approaches determine the location of change first and then estimate the size of changes. In addition, the penalties in ℓ_1 approaches are post tuned rather than pre-determined or specified. Several ℓ_1 methods, for example, ordinary LASSO and adaptive LASSO have been applied to the changepoint problems, but a comprehensive review of ℓ_1 methods and a detailed analysis of their performance have not been done, for example, which ℓ_1 method performs best, which information criterion best selects the regularization parameter, and are post-selections inevitable? If post-selections are needed, what post selection should be used?... I try to answer these questions in the doctorate research. When I was writing this dissertation, the related publication was still under preparation, so this chapter contains only the review of ℓ_1 regularization.

5.1 Linear Model of Changepoints and Ordinary LASSO

In this chapter we assume that $\epsilon_1, \epsilon_2, \cdots, \epsilon_N$ are independent identically distributed random variables with zero mean and variance σ^2 in the changepoint model

$$X_t = \begin{cases} \mu_0 + \epsilon_t, & \tau_0 \leq t < \tau_1, \\ \mu_1 + \epsilon_t, & \tau_1 \leq t < \tau_2, \\ \quad \vdots & \\ \mu_m + \epsilon_t, & \tau_m \leq t < \tau_{m+1}. \end{cases} \tag{5.1}$$

The mean shifts in a time series can be represented by a linear model. A changepoint occurs at the time $k(\ 2 \leq k \leq N)$ if and only if $\mu_k \neq \mu_{k-1}$, there we define a vector $\boldsymbol{\beta}$ of length N such that

$$\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_N)^T = (\mu_1, \mu_2 - \mu_1, ..., \mu_N - \mu_{N-1})^T, \tag{5.2}$$

where β_j equals to the mean shift size from time $j-1$ to j, $2 \leq j \leq N$, Therefore, $\boldsymbol{\beta}$ is the vector with each element except the first one that corresponds to a potential changepoint. The changepoint problem can be written as

$$\boldsymbol{X}_t = \boldsymbol{D}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{D}$ is an $N \times N$ design matrix such that the i^{th} row is $[\underbrace{1, \cdots, 1}_{i \text{ ones}}, \underbrace{0, \cdots, 0}_{(n-i) \text{ zeros}}]$, and $\boldsymbol{\epsilon}$ is the error vector. Since there are a few significant non-zero changepoints, the model is sparse. A straightforward thought is to determine the number and time of changes through the model selection by minimizing the objective function

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2} \left\| \boldsymbol{X} - \sum_{j=1}^{N} D_j \beta_j \right\|^2 + \lambda \sum_{j=2}^{N} \mathbb{1}(\beta_j \neq 0), \tag{5.3}$$

where $\mathbb{1}(\cdot)$ is the indicator function. We should note that the equation 5.1 is an nonconvex optimization which is infeasible to solve. [15] relaxed the ℓ_0 changepoint detection to the ℓ_1-norm form

which turns out to be the ordinary LASSO:

$$\hat{\beta}(lasso) = \arg\ \min_{\beta}\ \frac{1}{2}\left\|\boldsymbol{X} - \sum_{j=1}^{N} D_j\beta_j\right\|^2 + \lambda\sum_{j=2}^{N}|\beta_j|, \tag{5.4}$$

where λ is the regularization parameter which controls the model sparsity. The solution to the equation 5.1 relies on the choice of λ. However, it's of little avail to use K-fold cross validation on the changepoint problems in time series, because time series with changepoints occurred are intrinsically ordered, which hardly satisfies the random partition of data for doing K-fold cross validation. Therefore, Harchaoui[15] computed the regularization path for the equation 5.1 by Least Angle Regression(LARS) and then determined the solution through a rule of thumb. However, Harchaoui's work approach is controversial. First, Least Angle regression relates to the forward stepwise model selection: instead of entering the best variable at each step, the estimated parameters are increased in a direction equiangular to each one's correlations with the residual [10]. Weisberg has argued in the discussion section of [10] that Least Angle Regression is "especially sensitive to the effects of noise because LARS is based upon an iterative refitting of the residuals". Simulation studies also conclude that LARS is not an appropriate tool for obtaining the regularization path in changepoint analysis. Second, the selection criteria of [15] is the rule of thumb. The author failed to supplement sound theoretical proofs. Though he simulated the approach, the simulation was conducted on the data with higher signal-to-noise ratio.

To address the foregoing two issues, Coordinate Descent (CD) algorithm is first recommended in computing the regularization path. The idea behind CD is to optimize a objective function with respect to a single parameter at a time and then iterate through all parameters until the algorithm converges. CD is simpler to implement and faster for large-scale problems. In particular, finding changepoints in a sequence of length N involves N features. The simulation in my dissertation also suggests that CD outperforms LARS on various changepoint scenarios.

Next, Bayesian Information Criterion (BIC), Minimum Description Length (MDL) and modified Bayesian Information Criterion (mBIC) should be preferred over the rule of thumb in selecting the best penalty parameter from the regularization path. Zhang, Li & Tsai [50] have explored the regularization parameter selections using the generalized information criterion. He further showed that BIC can identify the model consistently and Akaike Information Criterion(AIC) tends to over-

fit with positive probability but it's asymptotically loss efficient. mBIC was developed by Zhang & Siegmund [49] to improve the performance of BIC in the context of change-point problems.

It's of interest to know which criterion can best identify the changepoints via ℓ_1-regularization. An initial simulation study suggests AIC always has a serious overfitting issue. It frequently identifies changepoints almost everywhere in a sequence having only a few number of changepoints. This result can be explained by its weak penalty on the number of changepoints, by placing changepoints everywhere it would minimize the $\log(\hat{\sigma}^2)$ thus to achieve the smallest value for the object function.

5.2 Other ℓ_1 Approaches on Changepoint Detection

The performance of ordinary LASSO have been thoroughly studied by the statistical society. Though the ordinary LASSO performs continuous shrinkage and avoids the major drawback of stepwise model selections, it fails in general to enjoy the oracle properties because it's biased and the bias does necessarily not diminish as $n \to \infty$. The disappointing fact that the ordinary LASSO is incompetent in changepoint detection has made us eye other ℓ_1-regularization methods.

5.2.1 Fused LASSO and Total Variation

Since a changepoint occurs at a location k if and only if $\beta_k \neq 0$, it's convenient to use the Dirac function δ which equals to 1 everywhere except point 0 at which the function value is 0, i.e. $\delta(\beta_k) = 1$ if and only if $\beta_k \neq 0$. If we assume there are at most k changepoints, then the following nonlinear programming (NLP) model will be established for the changepoint problem:

$$\begin{aligned} \min \quad & \|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{\beta}\|^2 \\ s.t. \quad & \sum_{i=2}^{N} \delta(\beta_i) \leq k \end{aligned} \tag{5.5}$$

NLP's are generally hard to solve but this particular problem can be solved in $O(N^2k)$ by dynamic programming. Be aware that the performance depends on k, i.e. the predetermined upper bound of the number of change-points. However, such upper bound may not always be available for real-life datasets. By picking the trivial upper bound N in the sense that every sample point can be a change-point, this method is then $O(N^3)$.

The difficulty of solving (5.5) is due to the non-continuity of Dirac functions δ. To get rid

of it, an alternative approach called total variation (TV) ([15]) model is introduced.

$$\min \quad \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{D\beta}\|^2 + \lambda \sum_{i=2}^{N} |\beta_i|, \tag{5.6}$$

where $\lambda > 0$ is a given constant. (5.6) is a LASSO model and can be solve in $O(n\log(n))$ independent from the number of changepoints. [4] generalized TV to multiple profiles changepoints detection problem. From now on in this section, assume $X, \beta \in \mathbb{R}^{N\times p}$ where p is the number of profiles and denote $\beta_{i,\bullet}$ to be the i-th row of β. The group fused LASSO is

$$\min \quad \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{D\beta}\|^2 + \lambda \sum_{i=1}^{N-1} \frac{\|\beta_{i,\bullet}\|}{r_i} \tag{5.7}$$

The regularity term penalizes the sum of Euclidean norms of β_i and thereby (5.7) is so-called joint TV model. Intuitively, λ determines how sensitive (5.7) is to variation of data. When λ is too large, this penalty will enforce many vector β_i to collapse to 0 and the solution may fail to capture some changepoints. If λ is too small, β_i is easily affected by the noise and the model suffers from overfitting issues. The $\{r_i\}_{i=1}^{N-1}$ are position-dependent weights which allow us to assign different penalties to distinct profiles. These weights are empirically chosen to be

$$r_i = \sqrt{\frac{n}{i(n-i)}}, \quad \forall i = 1, \ldots, N-1$$

One approach to solve (5.7) is by reformulating it as a group LASSO. Let $R \in \mathbb{R}^{N\times N-1}$ and $\alpha \in \mathbb{R}^{N\times p}$defined as follows

$$R_{i,j} = \begin{cases} r_j, & i \geq j \\ 0, & i < j \end{cases} \text{ and } \alpha_{i,\bullet} = \frac{\beta_{i,\bullet}}{r_i}, \ \forall i = 1, \ldots, N-1$$

Then $X = D\beta = R\alpha$ and (5.7) can be rewritten as a classical group LASSO model raised by [47].

$$\min \quad \frac{1}{2}\|\boldsymbol{X} - R\alpha\|^2 + \lambda \sum_{i=1}^{N-1} \|\alpha_{i,\bullet}\| \tag{5.8}$$

The group fused LASSO (5.7) is an extension of the total variation model to high dimensional data. Note that, suggested by the numerical results in [4], (5.7) is mainly used to detect approximately

shared change-points and its performance highly benefits from the number of profiles. For the sake of identifying changepoints in each individual profile, one needs to either apply post-selection or reformulate the model.

5.2.2 Oracle Property and Adaptive LASSO

Let $\mathcal{A} = \{j : \beta_j \neq 0\}$ be the set of coefficient indices of the true model and let $|A| = p_0$. Denote the coefficient estimator of a fitting procedure δ by by $\hat{\beta}(\delta)$. Fan and Li [11] has commented that a good model selection procedure should satisfy the following *properties*:

- it identifies the true model $\mathcal{A}$;
- it has the optimal estimation rate, i.e., $\sqrt{N}(\hat{\beta}(\delta)_{\mathcal{A}} - \beta_{\mathcal{A}}) \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma}^*)$, where $\mathbf{\Sigma}^*$ is the covariance matrix knowing the true model.

Alternative penalties are proposed to reduce the estimation bias of ordinary LASSO while maintaining the sparsity property. Zou [51] proposed the *adaptive LASSO* which is a two-stage least square approach, and proves to enjoy the oracle properties. Shen and Gallagher [40] applied the adaptive LASSO to the changepoint analysis on the climate data containing a linear trend. The adaptive LASSO identifies the changepoints via

$$\hat{\beta}(adalasso) = arg\min_{\beta} \left\| \boldsymbol{X} - \sum_{j=1}^{N} D_j \beta_j \right\|^2 + \lambda_n \sum_{j=2}^{N} \hat{w}_j |\beta_j|, \tag{5.9}$$

where λ_n is a nonnegative regularization parameter and $(\hat{w}_1, \cdots, \hat{w}_{N-1})$ is a weight vector for the changepoint parameters determined prior to the minimization. Zou [51] suggested any consistent initial estimator of $\boldsymbol{\beta}$ can be used as the weight vector, for example, the least squared estimates. However, they don't work for the changepoint problem because $N = p$. Alternative choice for the weight vector is ordinary LASSO estimates. Thus, the adaptive LASSO for finding changepoints is implemented as follows: we first run an ordinary LASSO to obtain the weight vector $(\hat{w}_2, \cdots, \hat{w}_N)$ with $\hat{w}_j = 1/|\hat{\beta}_j(lasso)|$; next we run the adaptive lasso to generate the final estimates for all $\beta_j' s$. In second stage of adaptive LASSO, the best penalty parameters λ_n is chosen via **BIC** and **mBIC**.

5.2.3 Non-convex Penalties

Different from the adaptive LASSO, another kind remedy is to use a penalty that tapers off as β increases in absolute value. Such an approach is single stage and the tapering penalty is no longer convex. **Smoothly Clipped Absolute Deviation** (SCAD) and . Both satisfy the oracle properties.

Fan and Li [11] proposed a non-concave penalty referred to as the **Smoothly Clipped Absolute Deviation** (SCAD). SCAD penalty is given by

$$P(\beta|a,\lambda) = \begin{cases} \lambda|\beta| & \text{if } |\beta| \leq \lambda; \\ \dfrac{|\beta|^2 - 2a\lambda|\beta| + \lambda^2}{2(a-1)} & \text{if } \lambda < |\beta| \leq a\lambda; \\ \dfrac{(a+1)\lambda^2}{2} & \text{if } |\beta| > a\lambda. \end{cases} \tag{5.10}$$

The solution to the SCAD penalty is given as

$$\hat{\beta}_j(SCAD) = \begin{cases} (|\hat{\beta}_j| - \lambda)_+ sign(\hat{\beta}_j) & \text{if } |\beta| \leq 2\lambda; \\ \dfrac{(a-1)\hat{\beta}_j - sign(\hat{\beta}_j)a\lambda}{a-2} & \text{if } 2\lambda < |\hat{\beta}_j| \leq a\lambda; \\ \hat{\beta}_j & \text{if } |\beta| > a\lambda. \end{cases} \tag{5.11}$$

The thresholding rule contains two tuning parameters λ and a. Theoretically, the best pair (λ, a) is obtained by two dimensional grid search using cross validation. However, the implementation is computational expensive. Fan and Li[11] suggested $a = 3.7$ as a default value for various problems from the perspective of Bayesian statistics and the results of simulation studies.

Another similar penalty, **Minimax Concave Penalty**(MCP) by Zhang [48] is of the form:

$$P(\beta|\gamma,\lambda) = \begin{cases} \lambda|\beta| - \dfrac{\beta^2}{2\gamma} & \text{if } |\beta| \leq \gamma\lambda; \\ \dfrac{1}{2}\gamma\lambda^2 & \text{if } |\beta| > \gamma\lambda. \end{cases} \tag{5.12}$$

Shen and Pan [41] proposed a surrogate to the ℓ_0 regularization problem defined by the equation 5.1. The ℓ_0 function is approximated by a truncated ℓ_1 function (TLP) which is defined as

$$P(\beta) = \min\left(\frac{|\beta|}{\tau}, 1\right), \tag{5.13}$$

where $\tau > 0$ is a tuning parameter that both controls the degree of approximation and decides which individual coefficients to be shrunk towards zero. The surrogate performs ℓ_0 model selection task while avoiding the computational infeasibility when minimizing a discontinuous cost function involving the ℓ_0 norm. In addition, the tuning parameter τ is able to discriminate small from large coefficient through thresholding.

Chapter 6

Conclusions and Discussion

6.1 Conclusions

The Sum of Squared CUSUMz has been shown to be the most powerful single changepoint technique but it cannot identify the changepoint time, so CUSUMz is needed to report the changepoint time when Sum of Squared CUSUMz test rejects the null hypothesis. Among the multiple changepoint techniques, it's hard to claim a winner and we know that the binary segmentation and MDL+GA should not be recommended for any changepoint analysis. Although BIC+GA generally performs better than others, the genetic algorithm cannot search for the changepoints from a long sequence due to the unacceptable computational cost. Moreover, we have developed an informative distance metric for changepoint comparisons and have tested its effectiveness in various simulation studies. It will be a popular distance metric in the changepoint research soon. Lastly, the Yule-Walker moment estimator based on the first order difference of the series can estimate autocorrelation parameters when the changepoints are relatively small. It's a key step to find changepoints correctly in a correlated time series though it has not been perfect yet, since it cannot handle frequent changepoints.

6.2 Theoretical Implications and Recommendations for Further Research

In chapter 3, we have discussed the weakness of the wild binary segmentation that it has a significantly high Type 1 error rate for a sequence of length 100 and without any changepoints. This is due to the inappropriate threshold used in [12]. Another issue comes from the random draws in the wild binary segmentation. It seems that these random draws are not as efficient as the author claimed. The short sub-intervals are less likely to contain a changepoint while the long sub-intervals are more likely to contain more than one changepoint in which the CUSUM test fails to detect. We appreciate its novelty, but wild binary segmentation needs to be reworked.

In chapter 4 we proposed a Yule-Walker moment estimator for the autocovariance estimation in the changepoint problems. The estimator is based on the first order difference of the series with the hope that the number of changepoints are much less than the length of the sequence. It's apparent that the estimator becomes more biased as the changepoint number increases. To correct the bias, we are developing a gradient descent PELT (GD-PELT) with Dr. Killick. The biased moment estimate is used as the initial value in the GD-PELT algorithm. GD-PELT algorithm will not only reduce the bias but also extend the i.i.d. based changepoint techniques to the AR(p) structure. In addition, the first order difference based moment estimator has broader applications. If a time series has both a linear trend and changepoints, the first order difference will eliminate the trend; if a time series has trend changes, second order difference can transform the trend change detection into a mean shift detection. These two applications are currently under our investigation.

ℓ_1-regularization has been reviewed in chapter 5. Though the simulation has not been completed, the initial result suggests it cannot compete against other changepoint techniques, which explains why a post selection is often inevitable. We are reconsidering a post selection using the genetic algorithm. The genetic algorithm is extremely slow when it has to search in a large solution domain. However, with the help of ℓ_1-regularization the domain will be significantly reduced, making the genetic algorithm perfectly suitable.

Besides the theories and methods, I'm also developing a software package to implement the work that I have done for the changepoint analysis. I will continue to work on these research projects during my postdoctoral time at the University of California.

Appendices

Appendix A Full Simulation Results of Single Changepoint Techniques

Table 1: Type I Error for Simulating AR(1) Series without changepoint. $\sigma^2 = 1$.

ϕ \ Test	CUSUM	λ_X	CUSUMz	λ_Z	SCUSUMz	W_T	LRT(W_U)	U_{crop}
$\phi = .9, N = 100$	0	0	0.0142	0.0341	0.046	**0.1053**	**0.0717**	**0.2261**
$(N = 500)$	(0.0084)	(0.0021)	(0.0279)	(0.0281)	(0.0409)	**(0.1143)**	**(0.0759)**	**(0.2898)**
$[N = 1000]$	[0.0179]	[0.0092]	[0.0336]	[0.0298]	[0.0448]	**[0.1193]**	**[0.0749]**	**[0.2755]**
$\{N = 2500\}$	{0.0308}	{0.0247}	{0.0435}	{0.0421}	{0.0512}	**{0.1216}**	**{0.0795}**	**{0.2369}**
$\phi = .8$	0.0007	0	0.0112	0.0152	0.0352	0.0281	0.0538	**0.1824**
	(0.0192)	(0.0096)	(0.032)	(0.029)	(0.0431)	(0.0344)	(0.0421)	**(0.1546)**
	[0.0313]	[0.0213]	[0.0423]	[0.0379]	[0.0479]	[0.039]	[0.0382]	**[0.1407]**
	{0.0367}	{0.0321}	{0.0431}	{0.042}	{0.0456}	{0.0392}	{0.0349}	**{0.1088}**
$\phi = .7$	0.0035	0.0003	0.0144	0.0148	0.0373	0.0107	0.0381	**0.1344**
	(0.0255)	(0.016)	(0.0387)	(0.0312)	(0.0468)	(0.0164)	(0.0283)	**(0.101)**
	[0.0316]	[0.0252]	[0.0398]	[0.0377]	[0.0467]	[0.0221]	[0.0274]	**[0.0901]**
	{0.043}	{0.0391}	{0.0478}	{0.0478}	{0.0477}	{0.022}	{0.0274}	**{0.0791}**
$\phi = .6$	0.0081	0.0018	0.0195	0.0119	0.0388	0.0041	0.0261	**0.1017**
	(0.0281)	(0.02)	(0.0353)	(0.0322)	(0.0459)	(0.0101)	(0.0204)	**(0.0744)**
	[0.0369]	[0.0312]	[0.0437]	[0.0402]	[0.0476]	[0.0125]	[0.0209]	**[0.0717]**
	{0.0445}	{0.0401}	{0.0497}	{0.0474}	{0.0519}	{0.0172}	{0.0235}	**{0.069}**
$\phi = .5$	0.0101	0.0045	0.0206	0.0128	0.0389	0.0038	0.0237	**0.0829**
	(0.0334)	(0.0285)	(0.0414)	(0.0387)	(0.0487)	(0.0091)	(0.0181)	**(0.0671)**
	[0.0341]	[0.0308]	[0.0388]	[0.038]	[0.0445]	[0.0104]	[0.0156]	**[0.0572]**
	{0.0426}	{0.0424}	{0.0455}	{0.0491}	{0.0489}	{0.0137}	{0.0171}	**{0.0643}**
$\phi = .4$	0.0148	0.0055	0.0209	0.0158	0.03407	0.0021	0.0185	**0.0667**
	(0.0314)	(0.0247)	(0.0364)	(0.0309)	(0.0437)	(0.0084)	(0.0139)	(0.0515)
	[0.0417]	[0.0381]	[0.0457]	[0.0423]	[0.051]	[0.0099]	[0.0155]	**[0.0558]**
	{0.0468}	{0.0426}	{0.0499}	{0.047}	{0.0533}	{0.014}	{0.0183}	**{0.0571}**
$\phi = .3$	0.018	0.0084	0.0244	0.0149	0.041	0.0025	0.0138	**0.0597**
	(0.0392)	(0.0335)	(0.0425)	(0.0392)	(0.0529)	(0.0103)	(0.0144)	(0.0542)
	[0.0466]	[0.0366]	[0.0486]	[0.0405]	[0.0518]	[0.0117]	[0.017]	[0.0511]
	{0.0436}	{0.0433}	{0.0455}	{0.0464}	{0.0482}	{0.0144}	{0.0164}	{0.053}
$\phi = .2$	0.0277	0.0117	0.0268	0.0177	0.0424	0.0024	0.0127	0.0507
	(0.0383)	(0.0343)	(0.0402)	(0.0378)	(0.049)	(0.0118)	(0.0153)	(0.0508)
	[0.0401]	[0.0396]	[0.0418]	[0.0428]	[0.0487]	[0.0112]	[0.0132]	[0.0495]
	{0.045}	{0.0425}	{0.0465}	{0.0438}	{0.047}	{0.0109}	{0.0116}	{0.0489}
$\phi = .1$	0.0218	0.0146	0.0238	0.017	0.0455	0.0029	0.0126	0.0436
	(0.0432)	(0.0364)	(0.0443)	(0.0383)	(0.0533)	(0.0086)	(0.0117)	(0.0465)
	[0.0419]	[0.0407]	[0.0431]	[0.0423]	[0.0449]	[0.013]	[0.0139]	[0.0459]
	{0.0459}	{0.044}	{0.0466}	{0.0449}	{0.0483}	{0.0153}	{0.0155}	{0.0474}
$\phi = -.25$	0.035	0.0321	0.0286	0.0223	0.045	0.0044	0.0102	0.0388
	(0.0407)	(0.0394)	(0.0376)	(0.0348)	(0.044)	(0.0084)	(0.0088)	(0.0386)
	[0.0478]	[0.0443]	[0.0453]	[0.041]	[0.0509]	[0.0117]	[0.0118]	[0.0408]
	{0.0441}	{0.0455}	{0.0423}	{0.0438}	{0.047}	{0.0121}	{0.0118}	{0.0435}
$\phi = -.5$	0.0466	0.0481	0.0306	0.0212	0.0464	0.005	0.0086	0.033
	(0.0519)	(0.0506)	(0.0432)	(0.0386)	(0.0504)	(0.0132)	(0.0118)	(0.0382)
	[0.05]	[0.0498]	[0.0449]	[0.0416]	[0.0498]	[0.011]	[0.0089]	[0.0407]
	{0.0513}	{0.0482}	{0.048}	{0.0445}	{0.0536}	{0.0135}	{0.0108}	{0.0426}
$\phi = -.9$	**0.1372**	**0.2928**	0.0295	0.0337	0.052	**0.1017**	0.007	0.0291
	(0.0845)	**(0.1096)**	(0.044)	(0.0423)	(0.0549)	(0.1228)	(0.0101)	(0.0364)
	[0.0705]	**[0.0856]**	[0.0451]	[0.0417]	[0.051]	**[0.1187]**	[0.0103]	[0.0388]
	{0.0595}	{0.0675}	{0.0452}	{0.0438}	{0.0528}	**{0.1161}**	{0.0093}	{0.0406}
$\phi = -.95$	**0.2698**	0.5276	0.0262	0.0429	0.0484	**0.2385**	0.0047	0.022
	(0.107)	**(0.1941)**	(0.0424)	(0.0469)	(0.0531)	**(0.2597)**	(0.0075)	(0.0359)
	[0.0925]	**[0.1331]**	[0.0485]	[0.0499]	[0.0535]	**[0.2644]**	[0.0107]	[0.0431]
	{0.0682}	**{0.0861}**	{0.0472}	{0.0447}	{0.0486}	**{0.2604}**	{0.0102}	{0.0412}

Table 2: Power. A changepoint at middle. $\sigma^2 = 1$, $\Delta = 0.15$.

Test / ϕ	CUSUM	λ_X	CUSUMz	λ_Z	SCUSSUM	T.LRT	LRT	cLRT
$\phi = .9, N = 100$	0	0	0.0151	0.0349	0.0487	0.1047	0.0729	0.2348
$(N = 500)$	(0.0096)	(0.0034)	(0.0318)	(0.027)	(0.0481)	(0.1149)	(0.0795)	(0.291)
$[N = 1000]$	[0.021]	[0.0109]	[0.0397]	[0.0338]	[0.0514]	[0.1129]	[0.0862]	[0.2853]
$\{N = 2500\}$	{0.0371}	{0.0238}	{0.0504}	{0.0417}	{0.0561}	{0.1172}	{0.0755}	{0.2403}
$\phi = .8$	0.0009	0.0001	0.0124	0.0167	0.0375	0.0311	0.0549	0.1809
	(0.0248)	(0.0116)	(0.0427)	(0.0328)	(0.0559)	(0.037)	(0.0478)	(0.1685)
	[0.0391]	[0.027]	[0.0555]	[0.0458]	[0.0673]	[0.0377]	[0.0479]	[0.1591]
	{0.08}	{0.0554}	{0.0927}	{0.0726}	{0.0963}	{0.0496}	{0.055}	{0.1592}
$\phi = .7$	0.0034	0.0003	0.0153	0.0131	0.0378	0.0119	0.0375	0.1271
	(0.039)	(0.0228)	(0.0547)	(0.0394)	(0.0681)	(0.0179)	(0.0354)	(0.1267)
	[0.0677]	[0.0442]	[0.0804]	[0.0596]	[0.0878]	[0.0234]	[0.0379]	[0.1285]
	{0.1443}	{0.0983}	{0.1568}	{0.1104}	{0.1707}	{0.0418}	{0.0609}	{0.1701}
$\phi = .6$	0.0084	0.0012	0.0218	0.0117	0.0439	0.0041	0.0312	0.108
	(0.0548)	(0.0315)	(0.0672)	(0.0472)	(0.0825)	(0.0157)	(0.031)	(0.1062)
	[0.0968]	[0.0643]	[0.1111]	[0.076]	[0.1281]	[0.0243]	[0.0428]	[0.1274]
	{0.2413}	{0.1668}	{0.2528}	{0.1786}	{0.2682}	{0.0597}	{0.0816}	{0.2284}
$\phi = .5$	0.014	0.0041	0.0233	0.0149	0.0462	0.0032	0.0252	0.0872
	(0.0765)	(0.0484)	(0.0879)	(0.0603)	(0.1072)	(0.0161)	(0.0337)	(0.109)
	[0.1479]	[0.0929]	[0.1601]	[0.1042]	[0.1762]	[0.03]	[0.0488]	[0.146]
	{0.3565}	{0.254}	{0.3676}	{0.2642}	{0.385}	{0.0967}	{0.1196}	{0.3061}
$\phi = .4$	0.0189	0.0064	0.0302	0.0155	0.0529	0.0028	0.0214	0.0834
	(0.1035)	(0.0651)	(0.1131)	(0.075)	(0.133)	(0.02)	(0.0368)	(0.1108)
	[0.2087]	[0.1347]	[0.2177]	[0.1463]	[0.2403]	[0.0446]	[0.0627]	[0.1835]
	{0.5003}	{0.3724}	{0.5078}	{0.3817}	{0.5173}	{0.165}	{0.1908}	{0.4216}
$\phi = .3$	0.0284	0.0132	0.0366	0.0221	0.0619	0.0039	0.0216	0.0808
	(0.1416)	(0.0869)	(0.1481)	(0.0952)	(0.1726)	(0.0275)	(0.0442)	(0.1319)
	[0.2813]	[0.1884]	[0.2894]	[0.1973]	[0.3096]	[0.0702]	[0.0931]	[0.234]
	{0.6351}	{0.4962}	{0.6418}	{0.5042}	{0.6468}	{0.2581}	{0.2904}	{0.5354}
$\phi = .2$	0.0315	0.0153	0.0385	0.0213	0.0646	0.0023	0.0198	0.069
	(0.1874)	(0.1178)	(0.1933)	(0.1244)	(0.2149)	(0.0363)	(0.0551)	(0.1588)
	[0.3597]	[0.2503]	[0.365]	[0.2569]	[0.3846]	[0.0982]	[0.1204]	[0.29]
	{0.7573}	{0.6374}	{0.7607}	{0.6409}	{0.7653}	{0.3795}	{0.4058}	{0.6655}
$\phi = .1$	0.0434	0.0249	0.0478	0.0276	0.0743	0.0039	0.02	0.0702
	(0.2324)	(0.1538)	(0.2352)	(0.1565)	(0.2614)	(0.0492)	(0.0687)	(0.1864)
	[0.4419]	[0.3219]	[0.4445]	[0.3242]	[0.4632]	[0.1343]	[0.1575]	[0.3545]
	{0.8574}	{0.7611}	{0.8578}	{0.7621}	{0.8577}	{0.5196}	{0.5468}	{0.7788}
$\phi = -.25$	0.0904	0.0566	0.0768	0.0436	0.1122	0.008	0.023	0.0792
	(0.4366)	(0.3113)	(0.4235)	(0.2964)	(0.4516)	(0.1227)	(0.1419)	(0.3241)
	[0.7493]	[0.6301]	[0.7426]	[0.6198]	[0.7544]	[0.3608]	[0.3876]	[0.6402]
	{0.9903}	{0.975}	{0.9899}	{0.9741}	{0.9887}	{0.9008}	{0.9105}	{0.9766}
$\phi = -.5$	0.1449	0.1026	0.1114	0.0628	0.1499	0.0135	0.0292	0.0942
	(0.6111)	(0.4775)	(0.5858)	(0.447)	(0.6091)	(0.217)	(0.2393)	(0.4654)
	[0.8967]	[0.8202]	[0.8897]	[0.8062]	[0.8902]	[0.5918]	[0.6078]	[0.8137]
	{0.9994}	{0.998}	{0.9994}	{0.9978}	{0.9993}	{0.9859}	{0.9867}	{0.998}
$\phi = -.9$	0.365	0.4273	0.1753	0.115	0.2329	0.1272	0.0441	0.1254
	(0.8563)	(0.7737)	(0.7981)	(0.6824)	(0.8088)	(0.5085)	(0.4626)	(0.6877)
	[0.9885]	[0.972]	[0.9827]	[0.9592]	[0.9811]	[0.8838]	[0.8732]	[0.9605]
	{1}	{1}	{1}	{1}	{1}	{0.9999}	{0.9999}	{1}
$\phi = -.95$	0.4785	0.6186	0.1748	0.1268	0.2368	0.2554	0.0425	0.1248
	(0.8958)	(0.8369)	(0.821)	(0.7123)	(0.8306)	(0.6087)	(0.4875)	(0.7184)
	[0.9922]	[0.9811]	[0.9855]	[0.9649]	[0.9851]	[0.9224]	[0.898]	[0.9673]
	{1}	{1}	{1}	{1}	{1}	{1}	{1}	{1}

Table 3: Power. A changepoint at middle. $\sigma^2 = 1$, $\Delta = 1$.

Test / ϕ	CUSUM	λ_X	CUSUMz	λ_Z	SCUSUM	T.LRT	LRT	cLRT
$\phi = .9, N = 100$	0	0	0.018	0.0312	0.0526	0.1026	0.0804	0.2584
$(N = 500)$	(0.022)	(0.0039)	(0.0703)	(0.0353)	(0.1043)	(0.1118)	(0.1662)	(0.4607)
$[N = 1000]$	[0.1164]	[0.0444]	[0.1795]	[0.0935]	[0.2093]	[0.1179]	[0.2569]	[0.5563]
$\{N = 2500\}$	{0.4578}	{0.2917}	{0.5102}	{0.3503}	{0.5377}	{0.2144}	{0.5076}	{0.7689}
$\phi = .8$	0.0012	0	0.0229	0.0133	0.0644	0.0238	0.1062	0.2939
	(0.2143)	(0.0852)	(0.297)	(0.1467)	(0.3598)	(0.0506)	(0.3501)	(0.6345)
	[0.622]	[0.4148]	[0.6762]	[0.4817]	[0.707]	[0.2063]	[0.6438]	[0.8661]
	{0.9862}	{0.9564}	{0.9884}	{0.9637}	{0.9891}	{0.8548}	{0.9827}	{0.9968}
$\phi = .7$	0.0072	0.0005	0.0419	0.0162	0.0957	0.0098	0.1397	0.3551
	(0.5939)	(0.3635)	(0.6597)	(0.4395)	(0.7009)	(0.1536)	(0.6802)	(0.8766)
	[0.96]	[0.8855]	[0.969]	[0.907]	[0.9692]	[0.6914]	[0.9637]	[0.9936]
	{1}	{1}	{1}	{1}	{1}	{0.9995}	{1}	{1}
$\phi = .6$	0.026	0.0023	0.0806	0.0193	0.157	0.0035	0.1995	0.4512
	(0.8994)	(0.7382)	(0.9209)	(0.781)	(0.9308)	(0.4615)	(0.9353)	(0.9869)
	[0.9994]	[0.9959]	[0.9997]	[0.9966]	[0.9992]	[0.9699]	[0.9997]	[1]
	{1}	{1}	{1}	{1}	{1}	{1}	{1}	{1}
$\phi = .5$	0.079	0.0099	0.1531	0.035	0.2607	0.003	0.3147	0.5849
	(0.9914)	(0.9515)	(0.9934)	(0.9626)	(0.9913)	(0.8067)	(0.9968)	(0.9998)
	[1]	[0.9999]	[1]	[0.9999]	[1]	[0.9997]	[1]	[1]
	{1}	{1}	{1}	{1}	{1}	{1}	{1}	{1}
$\phi = .4$	0.1655	0.0309	0.2514	0.0664	0.3742	0.0054	0.4492	0.7107
	(0.9995)	(0.9954)	(0.9995)	(0.9964)	(0.9994)	(0.9669)	(1)	(1)
	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]
	{1}	{1}	{1}	{1}	{1}	{1}	{1}	{1}
$\phi = .3$	0.2935	0.076	0.3797	0.1237	0.5075	0.0116	0.6175	0.8368
	(1)	(0.9999)	(1)	(0.9999)	(1)	(0.9977)	(1)	(1)
	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]
	{1}	{1}	{1}	{1}	{1}	{1}	{1}	{1}
$\phi = .2$	0.4692	0.1671	0.5345	0.2272	0.6584	0.0297	0.7698	0.9289
	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)
	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]
	{1}	{1}	{1}	{1}	{1}	{1}	{1}	{1}
$\phi = .1$	0.6592	0.3131	0.6986	0.3715	0.7882	0.0778	0.8914	0.9728
	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)
	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]
	{1}	{1}	{1}	{1}	{1}	{1}	{1}	{1}
$\phi = -.25$	0.9872	0.913	0.9858	0.9121	0.9911	0.6183	0.9991	1
	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)
	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]
	{1}	{1}	{1}	{1}	{1}	{1}	{1}	{1}
$\phi = -.5$	0.9997	0.9972	0.9998	0.997	0.9998	0.9569	0.9999	1
	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)
	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]
	{1}	{1}	{1}	{1}	{1}	{1}	{1}	{1}
$\phi = -.9$	1	1	1	1	1	1	1	1
	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)
	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]
	{1}	{1}	{1}	{1}	{1}	{1}	{1}	{1}
$\phi = -.95$	1	1	1	1	1	1	1	1
	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)
	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]
	{1}	{1}	{1}	{1}	{1}	{1}	{1}	{1}

Appendix B R Codes for Doctorate Research

In Appendix B, the major parts of R code for my doctorate research are attached for both readers' convenicen and peers' review.

B.1 Critical Values of Browning Bridges

```
## Simulate for the quantitle for Brownain Bridges
## Brownian Bridge CDF
bbcdf = function(x) {
  k = seq(1:200)
  1 + 2 * sum( (-1)^k * exp(- 2 * k^2 * x^2))
}
# bbcdf(1.358099)   ##95% Quantile

bbquantile <- function(p) {
  if (p <= 0.5)
    stop("don't use less than 50% confidence")
  foo = function(x) bbcdf(x) - p
  return (uniroot(foo, c(0.5, 10))$root)
  ##uniroot searches the interval
  ##from lower to upper for a root
}
# bbquantile(0.95)
```

B.2 Critical Values of Cropped Browning Bridges

```
## Simulate the pvalue for the quantile
## of cropped Brownian Bridge Dist.
```

```
## sup_{l<t<h} B^2(t)/t(1-t) > x

l = 0.05
h = 0.95
x = 0.05

adj_bbcdf = function(x,l,h){
  ## x = test statistic
  ## l = lower cropping
  ## h = upper cropping
  sqrt(x*exp(-x)/(2*pi))*( (1-1/x) * log10((1-l)*h/(l*(1-h))) + 4/x )
}
adj_bbcdf(1,l,h)



adj_bbquantile = function(p) {
  if (p <= 0.5)
    stop("don't use less than 50% confidence")
  foo = function(x) adj_bbcdf(x,l,h) - p
  return (uniroot(foo, c(0.5, 10))$root)
  ##uniroot searches the interval
  ##from lower to upper for a root
}
```

B.3 Critical Values of the Integral Squared Brownian Bridge

```
N=100; lambda=0.45

Dp = function(x){
  (exp(-z^2/4)/gamma(0.5) )*(integrate(integrand, 0, Inf)$value)
```

```
}


##Based on eqn(2)-See Tolmatz(2002)
Fcdf2 = function(lambda){
  #lambda = 0.6
  N = 500
  item = NULL
  for(k in 1:N){
    integrand = function(x){ exp(-0.5*lambda*x^2)/sqrt(-x*sin(x)) }
    item[k]= (integrate(integrand, (2*k-1)*pi, (2*k)*pi)$value)
  }
  1-(2/pi)*sum(item) - 0.99 #is used to search root
}

uniroot(Fcdf2,c(0,1), tol = 0.0001)
```

B.4 Using Binary Search of Genetic Algorithm to Solve MDL

```
MDL.bin = function(loc.ind, Xt=xt){
  ##S1: setup for MDL computing
  loc.ind[1]=0
  N = length(Xt) #length of the series
  m = sum(loc.ind) #Number of CPTs
  if(m==0){
    mu.hat = mean(Xt)
    phi.hat = sum((Xt-mu.hat)[-N]*(Xt-mu.hat)[-1])/sum((Xt-mu.hat)[-1]^2)
    Xt.hat = c(mu.hat, mu.hat+phi.hat*(Xt[-N]-mu.hat))
    sigma.hatsq = sum( (Xt-Xt.hat)^2 )/N
    MDL=0.5*N*log(sigma.hatsq)
```

```
  }
  else{
    tau.vec = loc.ind*(1:N) #compute CPT index
    tau = tau.vec[tau.vec>0] #keep CPT locations only
    tau.ext = c(1,tau,(N+1)) #include CPT boundary 1 and N+1

    ##S2: Split Xt to compute sigma.hat.sq and phi.hat
    seg.len = diff(tau.ext) #length of each segments
    ff = rep(0:m, times=seg.len) ##create factors for segmentation
    Xseg = split(Xt, ff) ##Segmentation list
    mu.seg = unlist(lapply(Xseg,mean), use.names=F)
    mu.hat = rep(mu.seg, seg.len)
    phi.hat = sum((Xt-mu.hat)[-N]*(Xt-mu.hat)[-1])/sum((Xt-mu.hat)[-1]^2)
    Xt.hat = c(mu.hat[1], mu.hat[-1]+phi.hat*(Xt[-N]-mu.hat[-N]))
    sigma.hatsq = sum( (Xt-Xt.hat)^2 )/N
    MDL = 0.5*N*log(sigma.hatsq)+sum(log(diff(tau.ext))/2)+log(m)+sum(log(tau[-1]))
  }
  return(-MDL)
}

#Run MDL-GA
MDLGA = GA::ga(type="binary", fitness = MDL.bin,
               nBits = N, maxiter = 30000, run = 3000,
               popSize = 200, monitor = F)
```

B.5 Compute the Minimum Distance between Two Changepoint Configurations

```
cpt.dist = function(C1, C2, N){
```

```
  m = length(C1)
  k = length(C2)

  ##Generate Cost Matrix via all paired distance
  pair = expand.grid(C1, C2)
  if(m==k){
    cost.mat = matrix(abs(pair[,1]-pair[,2]),
                      nrow=m,ncol=k,byrow=T )
  }else if(m > k){ #C1 has more changepoints than C2
    cost.mat = cbind(matrix(abs(pair[,1]-pair[,2]),
                            nrow=m,ncol=k,byrow=T ),
                     matrix(0, nrow=m, ncol=(m-k), byrow=T))
  }else{ #C1 has less changepoints than C2, nrow < ncol
    cost.mat = rbind(matrix(abs(pair[,1]-pair[,2]),
                            nrow=m,ncol=k,byrow=F ),
                     matrix(0, nrow=(k-m), ncol=k, byrow= T ))
  }
  cpt.asgn = lpSolve::lp.assign(cost.mat, direction = "min")

  return( cpt.asgn$objval/N + abs(m-k) )
}
```

Bibliography

[1] Alexander Aue and Lajos Horváth. Structural breaks in time series. *Journal of Time Series Analysis*, 34(1):1–16, 2013.

[2] R. Baranowski, Y. Chen, and P. Fryzlewicz. Narrowest-over-threshold detection of multiple change points and change-point-like features. *Journal of the Royal Statistical Society: Series B*, 81(3):649–672, 2019.

[3] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[4] Kevin Bleakley and Jean-Philippe Vert. The group fused lasso for multiple change-point detection. *arXiv preprint arXiv:1106.4199*, 2011.

[5] Peter J Brockwell, Richard A Davis, and Stephen E Fienberg. *Time series: theory and methods: theory and methods*. Springer Science & Business Media, 1991.

[6] Rainer Burkard, Mauro Dell'Amico, and Silvano Martello. *Assignment problems, revised reprint*, volume 106. Siam, 2012.

[7] Souhil Chakar, E Lebarbier, Céline Lévy-Leduc, Stéphane Robin, et al. A robust approach for estimating change-points in the mean of an AR(1) process. *Bernoulli*, 23(2):1408–1447, 2017.

[8] Miklos Csorgo and Lajos Horváth. *Limit theorems in change-point analysis*. John Wiley & Sons Chichester, 1997.

[9] Richard A Davis, Thomas C M Lee, and Gabriel A Rodriguez-Yam. Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, 101(473):223–239, 2006.

[10] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[11] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[12] Piotr Fryzlewicz et al. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281, 2014.

[13] Colin Gallagher, Robert Lund, and Michael Robbins. Changepoint detection in daily precipitation data. *Environmetrics*, 23(5):407–419, 2012.

[14] Zhenguo Gao, Zuofeng Shang, Pang Du, and John L Robertson. Variance change point detection under a smoothly-changing mean trend with application to liver procurement. *Journal of the American Statistical Association*, 114(526):773–781, 2019.

[15] Zaid Harchaoui and Céline Lévy-Leduc. Catching change-points with lasso. In *NIPS*, volume 617, page 624, 2007.

[16] Zaıd Harchaoui and Céline Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492):1480–1493, 2010.

[17] Kaylea Haynes, Idris A Eckley, and Paul Fearnhead. Computationally efficient changepoint detection for a range of penalties. *Journal of Computational and Graphical Statistics*, 26(1):134–143, 2017.

[18] Carla Inclan and George C Tiao. Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, 89(427):913–923, 1994.

[19] Venkata Jandhyala, Stergios Fotopoulos, Ian MacNeill, and Pengyu Liu. Inference for single and multiple change-points in time series. *Journal of Time Series Analysis*, 34(4):423–446, 2013.

[20] Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

[21] Claudia Kirch. *Resampling methods for the change analysis of dependent data*. PhD thesis, Universität zu Köln, 2006.

[22] Claudia Kirch and Birte Muhsal. A mosum procedure for the estimation of multiple random change points. *Preprint*, 2014.

[23] Shanghong Li and Robert Lund. Multiple changepoint detection via genetic algorithms. *Journal of Climate*, 25(2):674–686, 2012.

[24] Y. Li, R. B. Lund, and A. Hewaarachchi. Multiple changepoint detection with partial information on changepoint times. *Electronic Journal of Statistics*, 13(2):2462–2520, 2019.

[25] Qiqi Lu and Robert B Lund. Simple linear regression with multiple level shifts. *Canadian Journal of Statistics*, 35(3):447–458, 2007.

[26] Robert Lund and Xueheng Shi. Detecting possibly frequent change-points: Wild binary segmentation 2. *arXiv preprint arXiv:2006.10845*, 2020.

[27] Robert Lund, Xiaolan L Wang, Qi Qi Lu, Jaxk Reeves, Colin Gallagher, and Yang Feng. Changepoint detection in periodic and autocorrelated time series. *Journal of Climate*, 20(20):5178–5190, 2007.

[28] Louis Lyons et al. *Statistical Problems in Particle Physics, Astrophysics and Cosmology: PHYSTAT05, Oxford, UK, 12-15 September 2005*. Imperial College Press, 2006.

[29] Ian B MacNeill. Tests for change of parameter at unknown times and distributions of some related functionals on brownian motion. *The Annals of Statistics*, pages 950–962, 1974.

[30] Ian B MacNeill et al. Properties of sequences of partial sums of polynomial regression residuals with applications to tests for change of regression at unknown times. *The Annals of Statistics*, 6(2):422–433, 1978.

[31] J Murray Mitchell Jr. On the causes of instrumentally observed secular temperature trends. *Journal of Meteorology*, 10(4):244–261, 1953.

[32] Yue S Niu, Ning Hao, and Heping Zhang. Multiple change-point detection: A selective overview. *Statistical Science*, pages 611–623, 2016.

[33] ES Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–527, 1955.

[34] Sidney I Resnick. *Adventures in stochastic processes*. Springer Science & Business Media, 1992.

[35] Jorma Rissanen. *Stochastic complexity in statistical inquiry*. World Scientific, 1989.

[36] Michael Robbins, Colin Gallagher, Robert Lund, and Alexander Aue. Mean shift testing in correlated data. *Journal of Time Series Analysis*, 32(5):498–511, 2011.

[37] Michael W Robbins, Colin M Gallagher, and Robert B Lund. A general regression changepoint test for time series data. *Journal of the American Statistical Association*, 111(514):670–683, 2016.

[38] Andrew Jhon Scott and M Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, pages 507–512, 1974.

[39] Luca Scrucca et al. Ga: a package for genetic algorithms in r. *Journal of Statistical Software*, 53(4):1–37, 2013.

[40] Jie Shen, Colin M Gallagher, and QiQi Lu. Detection of multiple undocumented change-points using adaptive lasso. *Journal of Applied Statistics*, 41(6):1161–1173, 2014.

[41] Xiaotong Shen, Wei Pan, and Yunzhang Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497):223–232, 2012.

[42] X. Shi, C. M. Gallagher, R. B. Lund, and R. Killick. A statistical comparison of single and multiple changepoint techniques for time series data. *In Preparation*, 2020.

[43] Xueheng Shi and Colin Gallagher. Estimating unknown cycles in geophysical data. *Earth and Space Science*, 2019.

[44] Petre Stoica, Randolph L Moses, et al. *Spectral analysis of signals*. Pearson Prentice Hall Upper Saddle River, NJ, 2005.

[45] Leonid Tolmatz et al. On the distribution of the square integral of the brownian bridge. *The Annals of Probability*, 30(1):253–269, 2002.

[46] D. Wang, Y. Yu, and A. Rinaldo. Univariate mean change point detection: Penalization, CUSUM and optimality. *Electronic Journal of Statistics*, 14(1):1917–1961, 2020.

[47] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[48] Cun-Hui Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

[49] Nancy R Zhang and David O Siegmund. A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32, 2007.

[50] Yiyun Zhang, Runze Li, and Chih-Ling Tsai. Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105(489):312–323, 2010.

[51] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.